# CHEMOMETRICS-BASED SPECTROSCOPY FOR PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

EDITED BY: Vu Dang Hoang and Federico Marini
PUBLISHED IN: Frontiers in Chemistry

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-845-5
DOI 10.3389/978-2-88945-845-5

## About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

# CHEMOMETRICS-BASED SPECTROSCOPY FOR PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

Topic Editors:
**Vu Dang Hoang,** Hanoi University of Pharmacy, Vietnam
**Federico Marini,** Sapienza University of Rome, Italy

Chemometrics is the application of mathematics and statistics to chemical data in order to design or select optimal experimental procedures, to provide maximum relevant information, and to obtain knowledge about systems under study. This chemical discipline has constantly developed to become a mature field of Analytical Chemistry after its inception in the 1970s. The utility and versatility of chemometric techniques enable spectroscopists to perform multidimensional classification and/or calibration of spectral data that make identification and quantification of analytes in complex mixtures possible.

Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. They are now being adapted for a vast number of signal processing due to their unprecedented success in terms of asymptotic optimality, spatial adaptivity and computational efficiency. In analytical chemistry, they have increasingly shown great applicability and have been preferred over existing signal processing algorithms in noise removal, resolution enhancement, data compression and chemometrics modeling in chemical studies.

The aim of this Research Topic is to present state-of-the-art applications of chemometrics, in the field of spectroscopy, with special attention to the use of wavelet transform. Both reviews and original research articles on pharmaceutical and biomedical analysis are welcome in the specialty section Analytical Chemistry.

**Citation:** Hoang, V. D., Marini, F., eds. (2019). Chemometrics-based Spectroscopy for Pharmaceutical and Biomedical Analysis. Lausanne: Frontiers Media.
doi: 10.3389/978-2-88945-845-5

# Table of Contents

frontiers in Chemistry

EDITORIAL
published: 27 March 2019
doi: 10.3389/fchem.2019.00153

# Editorial: Chemometrics-based Spectroscopy for Pharmaceutical and Biomedical Analysis

*Hoang Vu Dang*[1]* *and Federico Marini*[2]

[1] Department of Analytical Chemistry and Toxicology, Hanoi University of Pharmacy, Hanoi, Vietnam, [2] Department of Chemistry, Sapienza University of Rome, Rome, Italy

**Keywords: chemometrics, spectroscopy, Pharmaceutical analysis, biomedical analysis, Wavelet Transform**

## Editorial on the Research Topic

## Chemometrics-based Spectroscopy for Pharmaceutical and Biomedical Analysis

Spectroscopy is associated with a plethora of different techniques studying the interaction between matter and electromagnetic radiation. Linguistically speaking, the term originates from the Latin word "spectrum" meaning "specter or image/vision," and the Greek word "σκοπεῖν" meaning "to view or inspect." In other words, it is concerned with the absorption, emission, or scattering of electromagnetic radiation of different wavelengths, intimately linked to the structure of atoms or molecules under study.

Unambiguously, spectroscopy and optical measurement technologies are of great importance for analysis of chemical composition. Spectroscopic techniques such as UV-Vis and IR are routinely used in laboratories as well as detailed in a great number of pharmacopeia monographs (e.g., United State pharmacopeia, British pharmacopeia and European pharmacopeia) for quality control of excipients, pharmaceutical ingredients and dosage forms. These techniques can offer a rapid, cheap, non-invasive/non-destructive analysis, using both off-line and in-/at-/on-line methodologies. Nevertheless, they are usually limited to the identification and assay by spectral comparison of a test sample against a reference standard. This approach may not be suitably applied to qualitative and quantitative analysis of real-world samples due to the complexity of pharmaceutical and biomedical matrices.

Given the above information, the use of chemometrics in spectroscopy is a must to gain efficiency in accessing spectral data. By definition, chemometrics is the use of mathematical and statistical methods to extract relevant chemical information and to correlate quality parameters or physical properties to analytical data. It means that a chemometrician would refer to the knowledge of chemical and instrumental influences to display in ways allowing chemical interpretation of the system under study (Davies, 2012).

With reference to the most straight-forward explanation of chemometrics, in the present Research Topic, Biancolillo and Marini briefly reviewed the different chemometric approaches applicable in the context of spectroscopy-based pharmaceutical analysis, discussing the unsupervised exploration of the collected data as well as the possibility of building predictive models for both quantitative (calibration) and qualitative (classification) responses.

In another review, Tsenkova et al. described the up-to-date development of multivariate analysis methodology in aquaphotomics, a novel scientific discipline proposed by Tsenkova (2005). In aquaphotomics analysis, an aquaphotome (i.e., a database of water absorbance bands and patterns correlating water structures to their specific functions) is built by using light-water interaction. To deal with such complex multidimensional spectral data, chemometric methods are exploited to remove unwanted influences and extract water absorbance spectral patterns related to the perturbation of interest.

OPEN ACCESS

***Edited and reviewed by:***
*Huan-Tsung Chang,*
*National Taiwan University, Taiwan*

***Correspondence:**
*Hoang Vu Dang*
*hoangvd@hup.edu.vn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *30 November 2018*
***Accepted:*** *01 March 2019*
***Published:*** *27 March 2019*

***Citation:***
*Vu Dang H and Marini F (2019) Editorial: Chemometrics-based Spectroscopy for Pharmaceutical and Biomedical Analysis. Front. Chem. 7:153. doi: 10.3389/fchem.2019.00153*

In spectral analysis, wavelets have increasingly shown great potential in chemical studies by being superior to existing signal processing algorithms in noise removal, resolution enhancement, data compression, and chemometric modeling (Chau et al., 2004; Vu Dang, 2014). In practice, multicomponent analysis may not be possible with a traditional UV spectrophotometric method due to spectral overlapping of both active and inactive ingredients of pharmaceutical samples. Majorly based on a series of studies by Dinç and co-workers, the review of Dinç and Yazan clearly detailed the theoretical aspects of wavelet transform (i.e., discrete, continuous, and fractional) and its characteristic application to UV spectroscopic analysis of pharmaceuticals.

For pharmaceutical and biomedical analysis, it is noteworthy that the combination of various spectroscopic techniques is advisable in an effort to scrutinize a complex chemical process. In the present Research Topic, this is truly reflected by the following works: (i) Wani et al. studying interaction of neratinib (an anticancer drug) with bovine serum albumin by using both spectroscopic (spectrofluorometric, UV spectrophotometric and Fourier-transform infrared) and molecular docking approaches, and (ii) Shang et al. designing and synthesizing low-cytotoxicity fluorescent probes based on anthracene derivatives for hydrogen sulfide detection.

Nowadays, the on-going application of vibrational spectroscopy has been increasingly generating an enormous number of papers published in the pharmaceutical and biomedical sciences (Abramczyk et al., 2017; Brody et al., 2017; Bunaciu and Aboul-Enein, 2017). It is thus not surprising that the present Research Topic mainly consists of research articles related to infrared and Raman spectroscopy. For instance, Tian et al. explored the use of chemometrics-based Fourier transform infrared spectroscopy for the investigation of plasma biochemical changes due to acute lead poisoning in a rat model. Ryabchykov et al. investigated a data fusion approach for combining the two most powerful imaging techniques (Raman spectroscopy and matrix-assisted laser desorption/ionization mass spectrometry) to better distinguish different regions within biological samples. Risoluti and Materazzi coupled a miniaturized Near Infrared (NIR) spectrometer to chemometrics as a novel entirely on-site approach for assessment of occupational exposure to hydroxyurea. Zou et al. compiled a NIR spectral library of amoxicillin and potassium clavulanate by using a universal model to resolve sample-collection problems, making quantitative models more specific for Process Analytical Technology control. Dai et al. discovered the linear region of Near Infrared Diffuse Reflectance spectra of different particle sizes by using the Kubelka-Munk theory, to serve as a methodological reference for the performance of prediction models. Chen et al. introduced a novel strategy for the real-time quantification of potassium in infant formula samples, i.e., applying a modified random frog algorithm, adopted in a higher-density discrete wavelet transform domain, to select the most important features of laser-induced breakdown spectra related to potassium. Zhao et al. proved that a pharmaceutical analysis model could be more reliable and robust when its parameters (such as spectral pretreatment, latent factors, variable selection, and calibration methods) were optimized by processing trajectory, possibly integrated into PLS software. Bogomolov et al. suggested a time-domain averaging of spectral variables to improve the accuracy of in-line NIR spectroscopic moisture monitoring in a fluidized bed drying process of pharmaceutical powder. Ma et al. proposed the use of the low-rank estimation method to improve the accuracy and robustness of Partial Least Squares and Support Vector Machine chemometric models being applied to Raman quantitative analysis of pharmaceutical mixtures.

Regarding the instrumentation for vibrational spectroscopy, Chen et al. developed a moving window fast Fourier transform cross-correlation to correct non-linear shifts for synchronization of spectra obtained from different Raman instruments. In another study, Fujiwara and Kano recommended the nearest correlation—based input variable weighting method for efficient and highly-accurate soft-sensor design, which is applicable to NIR data especially when the number of input variables is large.

The idea for this Research Topic originally came from the fact that the state-of-the-art application of chemometrics, in particular wavelet transform, plays a vital role in the field of spectroscopy being unceasingly perfected and matured.

As the title indicates, hopefully, it will serve as a useful guide for spectroscopic analysis in the pharmaceutical and biomedical sciences.

## AUTHOR CONTRIBUTIONS

HV wrote and FM revised the manuscript.

## REFERENCES


Abramczyk, H., Kopec, M., and Jedrzejczyk, M. (2017). "Raman spectroscopy, Medical applications: A new look inside human body with Raman imaging," in *Encyclopedia of Spectroscopy and Spectrometry, 3rd Edn.*, eds J. C. Lindon, G. E. Tranter, and D. W. Koppenaal (Cambridge, MA: Academic Press), 915–918.

Brody, R. H., Carter, E. A., Edwards, H. G. M., and Pollard, A. M. (2017). "FT-Raman spectroscopy applications," in *Encyclopedia of Spectroscopy and Spectrometry, 3rd Edn.*, eds J. C. Lindon, G. E. Tranter, and D. W. Koppenaal (Cambridge, MA: Academic Press), 770–777.

Bunaciu, A. A., and Aboul-Enein, H. Y. (2017). "Vibrational spectroscopy applications in drugs analysis," in *Encyclopedia of Spectroscopy and Spectrometry, 3rd Edn.*, eds J. C. Lindon, G. E. Tranter, and D. W. Koppenaal (Cambridge, MA: Academic Press), 575–581.

Chau, F. T., Liang, Y. Z., Gao, J., and Shao, X. G. (2004). *Chemometrics from Basics to Wavelet Transform*. Hoboken, NJ: John Wiley and Sons, Inc.

Davies, A.M. C. (2012). What IS and what is NOT chemometrics. *Eur. Spectrosc.* 24, 33–36.

Tsenkova, R. (2005). "Visible-near infrared perturbation spectroscopy: water in action seen as a source of information," in *12th International Conference on Near-infrared Spectroscopy* (Auckland), 607–612.

Vu Dang, H. (2014). Wavelet-based spectral analysis. *TRAC-Trend Anal. Chem.* 62, 144–153. doi: 10.1016/j.trac.2014.07.010


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 07 March 2018
doi: 10.3389/fchem.2018.00047

# Study of Interactions of an Anticancer Drug Neratinib With Bovine Serum Albumin: Spectroscopic and Molecular Docking Approach

*Tanveer A. Wani*[1]*, *Ahmed H. Bakheit*[1,2], *M. A. Abounassif*[1] and *Seema Zargar*[3]

[1] Department of Pharmaceutical Chemistry, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia, [2] Department of Chemistry, Faculty of Science and Technology, Al-Neelain University, Khartoum, Sudan, [3] Department of Biochemistry, College of Science, King Saud University, Riyadh, Saudi Arabia

Binding of therapeutic agents to plasma proteins, particularly to serum albumin, provides valuable information in the drug development. This study was designed to evaluate the binding interaction of neratinib with bovine serum albumin (BSA). Neratinib blocks HER2 signaling and is effective in trastuzumab-resistant breast cancer treatment. Spectrofluorometric, UV spectrophotometric, and fourier transform infrared (FT-IR) and molecular docking experiments were performed to study this interaction. The fluorescence of BSA is attributed to the presence of tryptophan (Trp) residues. The fluorescence of BSA in presence of neratinib was studied using the excitation wavelength of 280 nm and the emission was measured at 300-500 nm at three different temperatures. Neratinib quenched the BSA intrinsic fluorescence by static mechanism. A complex formation occurred due to the interaction leading to BSA absorption shift. The fluorescence, UV- absorption, three dimensional fluorescence and FT-IR data showed conformational changes occurred in BSA after interaction with neratinib. The binding constant values decreased as the temperature increased suggesting an instable complex formation at high temperature. Site I (sub-domain IIA) was observed as the principal binding site for neratinib. Hydrogen bonding and Van der Waals forces were suggested to be involved in the BSA-neratinib interaction due to the negative values of entropy and enthalpy changes.

**Keywords: bovine serum albumin, neratinib, human serum albumin, fluorescence, quenching**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Hui Xu,*
*Ludong University, China*
*Simone Brogi,*
*University of Siena, Italy*

***Correspondence:**
*Tanveer A. Wani*
*twani@ksu.edu.sa*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *24 October 2017*
***Accepted:*** *22 February 2018*
***Published:*** *07 March 2018*

***Citation:***
*Wani TA, Bakheit AH, Abounassif MA and Zargar S (2018) Study of Interactions of an Anticancer Drug Neratinib With Bovine Serum Albumin: Spectroscopic and Molecular Docking Approach. Front. Chem. 6:47. doi: 10.3389/fchem.2018.00047*

## INTRODUCTION

Neratinib, a tyrosine kinase inhibitor, is used in trastuzumab-resistant breast cancer treatment as an alternative to block HER2 signaling (**Figure 1**; Burstein et al., 2010; Iqbal and Iqbal, 2014; Wani et al., 2015). Neratinib has been recently approved by United States FDA for use in early stage HER2-overexpressed/amplified breast cancer (Bose and Ozer, 2009; Feldinger and Kong, 2015; Kourie et al., 2016; US Food and Drug Administration, 2018).

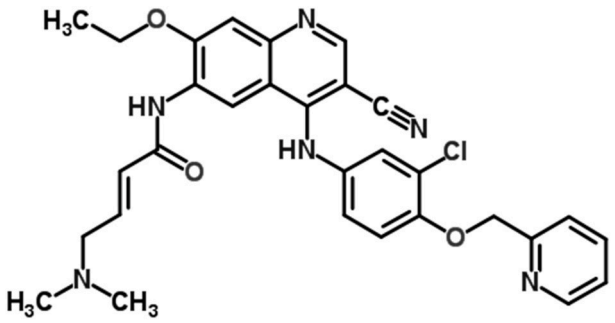


**FIGURE 1 |** Chemical structure of neratinib.

Plasma proteins act as carriers for transportation of drugs and other compounds. Amongst the various plasma proteins, serum albumin is the most abundant protein and it plays a vital role in transportation of drug ligands (Jahanban-Esfahlan et al., 2015; Wani et al., 2017b,c). Several tyrosine kinase inhibitors have been studied for their interaction with bovine serum protein (BSA) (Shen et al., 2015) and in this study, the interaction of neratinib with BSA was explored. BSA was selected for studying the interaction owing to its structural similarity to human serum albumin (HSA), low procurement cost and ready availability (He and Carter, 1992; Chi et al., 2010). So far, studies on the interaction between plasma proteins and neratinib only focused on the characterization of neratinib covalent binding with serum albumin and reversible covalent binding of neratinib with plasma proteins (Chandrasekaran et al., 2010; Wang et al., 2010). The BSA contains 583 amino acids and three homologous domains. These homologous I, II, and III domains are connected by disulfide bonds. Two tryptophan residues namely Trp-134 and Trp-212, are present in BSA molecule and have intrinsic fluorescence (Kragh-Hansen, 1981). The pharmacokinetics parameters of distribution, transportation and excretion of small ligands depend on the noncovalent binding interactions of drug ligands to proteins. Exploration of the interaction mechanism between the drug ligands with BSA is of great interest (Berezhkovskiy, 2007; Chamani and Heshmati, 2008; Xiao et al., 2011; Khorsand Ahmadi et al., 2015; Marouzi et al., 2017).

The interaction between neratinib and serum albumin was explored in this study. Multispectroscopic (UV-vis absorption, fluorescence, FT-IR) along with computational approaches were used to study the binding interaction. The parameters under study included binding site involvement, complex formation and binding energies of neratinib with BSA. The molecular docking data were corroborated with experimental results to obtain a better understanding of the mechanisms involved in the interaction.

# METHODS

## Chemicals and Reagents

Bovine serum albumin (BSA) was procured from Sisco Research Laboratories, India. Neratinib was obtained from Selleckchem, USA. Phenylbutazone and ibuprofen were purchased through National Scientific Company, Saudi Arabia. The stock solutions for neratinib, BSA, phenylbutazone and ibuprofen were prepared as per their molecular weight. Phosphate buffer pH 7.4 was used for preparation of BSA stock solution of $1.5 \times 10^{-6}$ M. Neratinib was dissolved in 500 μL dimethyl sulphoxide and then diluted with phosphate buffer pH 7.4 to get a stock concentration of $1.8 \times 10^{-3}$ M. The stock concentration was further diluted with the buffer to obtain working standard solutions in the range of $3.8 \times 10^{-5}$ and $5.2 \times 10^{-4}$ M. The stock solutions of ibuprofen and phenylbutazone were prepared in methanol and then diluted with the phosphate buffer. The deionized water was obtained from a Flex Type-IV instrument from Elga Lab Water, UK.

## Fluorescence Spectra Measurement

The fluorescence analysis was carried out using a JASCO FP-8200 spectrofluorometer (Japan). The chosen excitation wavelength was 280 nm and the emission fluorescence was attained within the 300–500 nm range. BSA solution $1.5 \times 10^{-6}$ M was titrated with different neratinib concentrations ($0, 1.5 \times 10^{-6}, \ldots, 2.11 \times 10^{-5}$ M) and the fluorescence measurements were carried out at the temperatures of 298, 303, and 308 K. These two solutions were mixed in a ratio of 1:1 *v/v*. Thus, the concentrations measured were half of the initial concentrations of either BSA or neratinib. The fluorescence intensity (FI) might decrease due to inner filter effect since a compound present in the solution might absorb in the ultraviolet region near the excitation/emission wavelength. Therefore, the correction of FI was done for studying the neratinib–BSA interaction using the following equation:

$$Fcor = Fobs \times e^{(Aex+Aem)/2}$$

Where, *Fcor* and *Fobs* denote corrected fluorescence intensity and measured fluorescence intensity respectively; and *Aex* and *Aem* are the modified absorbance values of the protein upon ligand addition at the excitation and emission wavelengths, respectively.

## Synchronous Fluorescence Spectra Measurement

The synchronous fluorescence spectra were studied for conformational changes that could occur in BSA at 298 K (room temperature). Scanning intervals Δλ (Δλ = λem-λex) of 15 and 60 nm characterize the tyrosine and tryptophan residues, respectively.

## FT-IR Spectra Measurement

A Bruker Alpha II FT-IR spectrometer (USA) coupled with the OPUS software was used. The spectra (spectral resolution 2 $cm^{-1}$; 24 scans) obtained were converted into absorbance. The spectra for the buffer and BSA solution in buffer were obtained, and the spectrum of buffer solution was subtracted from the BSA solution to get FT-IR spectra of BSA. Similarly, the BSA-neratinib solution was prepared and the spectra for the free neratinib was subtracted from the bound form. The FT-IR results provided evidence of possible conformational changes in the protein molecule.

## Site Probe Experiment

Site probe experiments were also conducted to determine the binding site involved in the interaction. Different concentrations of neratinib were added to equimolar concentrations of site probes (phenylbutazone or ibuprofen) and BSA; the FI was then determined at room temperature (298 K) and excitation wavelength of 280 nm.

## UV–Visible Spectra Measurement

The UV-Visible absorption spectra were attained in the range of 200–400 nm for BSA, neratinib and BSA-neratinib complex at room temperature (298 K) with a UV-1800 spectrophotometer (Shimadzu, Japan). The BSA-neratinib spectra were acquired by keeping BSA concentration constant (1.5 μM) and varying neratinib concentration.

## Molecular Docking

Molecular docking analysis was performed to studythe interaction between neratinib and BSA. The docking was performed on Molecular Operating Environment (MOE-2014). The structure for neratinib was drawn in the MOE, whereas the BSA crystalline protein structure was obtained from protein data bank (pdb) with the pdb code number 4OR0 (http://www.rcsb.org). Chain A of the BSA molecule was selected for the docking analysis due to the fact that BSA exist as a homodimer of two chains. Both protein receptors and ligands were protonated when prepared; and the energy minimization was performed with the default parameters of Force field MMFF94X, eps = r and cut off (8–10). The docking parameters used in the analysis were kept as default with Triangle Matcher. The rescoring function 1 was set as London dG and the rescoring function 2 was set as GBVI/WSA dG along with 10 conformation generations in order to fit the binding groove. mdb output file was generated for further analysis and evaluation of neratinib–BSA interaction. The active binding site that might be involved in the interaction was obtained from the site specific probe experiments (Jahanban-Esfahlan et al., 2015; Wani et al., 2017b,c). RMSD (root mean square deviation) parameters were used to select the most suitable interaction of BSA with neratinib.

# RESULTS

## Fluorescence Quenching

The FI of BSA and BSA-neratinib complex were recorded with excitation at 280 nm and emission in the range of 300–500 nm. The BSA concentration was kept constant whereas, the concentration of neratinib was varied. A decrease in FI was observed with increasing neratinib concentration. This was attributed to the quenching of fluorescence by BSA because of the formation of a non-fluorescent complex between neratinib and BSA (**Figure 2**). The quenching data was analyzed using the Stern-Volmer equation:

$$\frac{F}{F_0} = 1 + K_{sv}\,[Q] = 1 + K_q\,\tau_0\,[Q]$$

$F_0$ and $F$ represent the FIs in absence and presence of neratinib; K*sv:* Stern-Volmer quenching constant; [*Q*]: quencher concentration; $K_q$: quenching rate constant; $\tau_0$: fluorophore's lifetime devoid of quencher and is valued $10^{-8}$ for a biopolymer (Lakowicz and Weber, 1973). The values obtained for Ksv at the three different temperatures are presented in **Table 1** (**Figure 3**). During the synchronous fluorescence experiments, a stronger quenching of FI was observed for tryptophan residues $\Delta\lambda = 60$ nm compared to tyrosine residues $\Delta\lambda = 15$ nm indicating the contribution of tryptophan in the intrinsic fluorescence of BSA (**Figure 4**). Also a red shift equal to 1 nm was observed for tryptophan residue. The 3D (3-dimensional) spectrofluorometric analysis of BSA and BSA-neratinib complex (**Figure 5**) was performed indicating changes in the BSA conformation after addition of neratinib.

## Binding Constant

Small drug ligands interact with proteins binding sites independently and the equilibrium among the free and bound molecules is represented by the following equation (He et al., 2010):

$$\log\frac{(F_0 - F)}{F} = n\log K_b \pm n\log\left[\frac{1}{[Q] - \frac{(F_0 - F)[P]}{F_0}}\right]$$

Where $K_b$ is binding constant and n is binding site number; [*Q*] and [*P*] are the total concentrations of quencher and protein. A plot between log $(F_0\text{-}F)/F$ vs. log $\{1/([Q]\text{–}(F_0\text{-}F)\ [P]/F_0)\}$ is used to calculate the binding constant (intercept) and number of binding sites (slope). The binding constants and number of binding sites were determined at all the three temperatures and

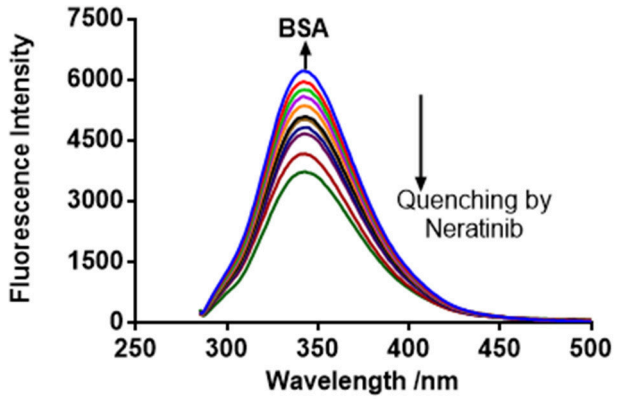



**FIGURE 2 |** BSA fluorescence quenching spectra in presence of neratinib at 298 K, λex = 280 nm.

**TABLE 1 |** Stern–Volmer quenching constants ($K_{SV}$) and bimolecular quenching rate constant (Kq) for the binding of neratinib to BSA at three different temperatures.

| T(K) | R | Ksv ± SD × $10^4$ (L $mol^{-1}$) | Kq × $10^{12}$ (L $mol^{-1}s^{-1}$) |
|---|---|---|---|
| 298 | 0.9918 | 6.54 ± 0.31 | 6.54 |
| 303 | 0.9907 | 6.28 ± 0.18 | 6.28 |
| 308 | 0.9935 | 5.96 ± 0.37 | 5.96 |

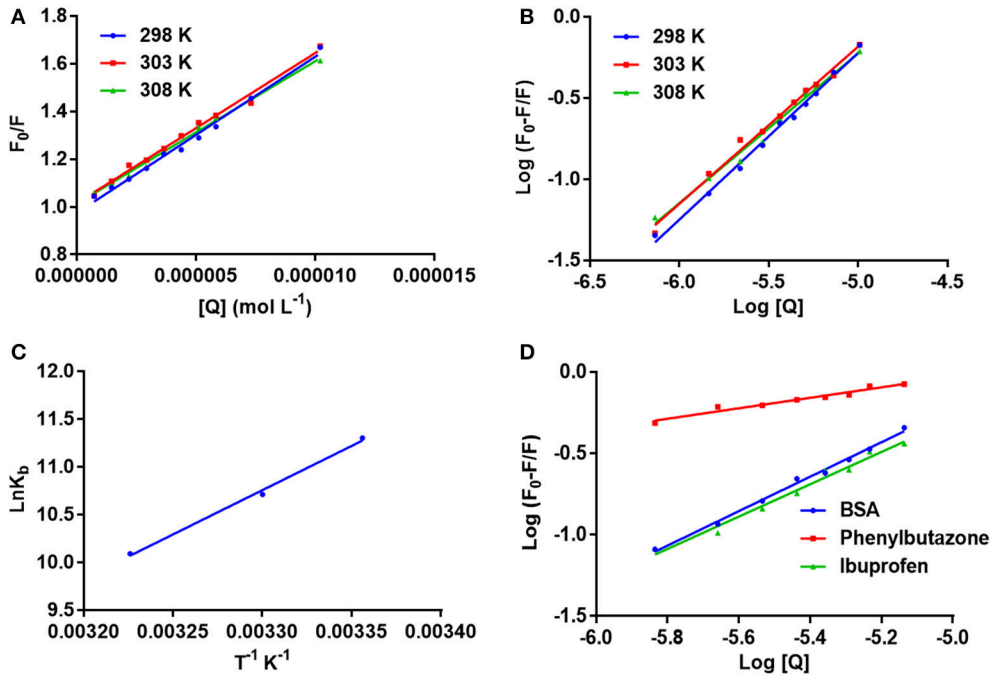



**FIGURE 3 | (A)** The stern–Volmer curves for the quenching of BSA by neratinib at 298/303/308 K; **(B)** The plot of log[($F_0$-F)/F] vs. log[Q] for quenching process of neratinib with BSA at 298/303/308 K; **(C)** Van't Hoff plots for the binding interaction of neratinib with BSA; **(D)** The plot of log[($F_0$-F)/F] vs. log[Q] for quenching process of neratinib with BSA in presence of site markers phenylbutazone and ibuprofen at 298 K.

are presented in **Table 2**. The number of binding sites were found equal to unity. The binding constant obtained for BSA-neratinib complex was found to be $8.1 \times 10^4$, whereas, in presence of phenylbutazone and ibuprofen were found to be $0.38 \times 10^2$ and $4.8 \times 10^4$, respectively (**Figure 3**).

## Binding Mode

The binding mode is established based on the thermodynamic parameters that include enthalpy change ($\Delta H^0$), entropy change ($\Delta S^0$) and free energy change ($\Delta G^0$). The thermodynamic parameters are given in **Table 2**. **Figure 3** represents the van't Hoff plot for neratinib and BSA interaction.

# DISCUSSION

## Neratinib Binding to the Serum Albumins

Fluorescence spectroscopy acts as a tool for investigation of the interaction between biological macromolecules (proteins) and small drug ligands. The interaction can be studied in terms of the mechanism involved in binding interaction, binding constants, etc. The FI can get reduced due to several molecular interactions that may include excited-state reactions, complex formations, energy transfer and molecular rearrangements. This decrease in the FI is known as fluorescence quenching. The type of quenching involved (static or dynamic) is derived from the linearity of the Stern-Volmer plot between $F_0$/F vs. [Q] (**Figure 3**). The Stern-Volmer plot alone cannot give sufficient information about the nature of quenching involved in the interaction. Thus, other evidences are still required for its determination. The change in temperature is used as a tool to investigate and distinguish between the static and dynamic quenching that may be involved in ligand-BSA interaction. The Ksv value decreases at higher temperature in static quenching, and vice versa in case of dynamic quenching. These results infer that a static quenching and complex formation could occur between neratinib and BSA. It was further supported by the quenching rate constants obtained (**Table 1**). The quenching constant for collision quenching can achieve a maximum value $2 \times 10^{10}$ $M^{-1}$ $S^{-1}$ for biopolymers. Our quenching constant values were much higher than those obtained by scattered procedure clearly showing the involvement of static quenching in the BSA-neratinib interaction (Shi et al., 2014; Wani et al., 2017a).

The synchronous fluorescence spectrophotometric experiments were performed to obtain information regarding the microenvironment present in the immediate neighborhood of chromosphere molecules. The conformational changes were reflected by changes in the maximum emission wavelength. A higher quenching and red shift of 1 nm was observed for tryptophan residue suggesting an increase in polarity of the surrounding environment (**Figure 4**). Therefore, it was concluded that the BSA conformation changes upon interaction of neratinib with BSA (Albert et al., 2006; Meti et al., 2015).

In the 3-dimensional spectral analysis for BSA in presence of neratinib, two peaks were found namely Peak 1 and Peak 2 (**Figure 5**). Peak 1 was found at the excitation wavelength of 230 nm and emission wavelength of 344 nm. Peak 1 is formed due to π-π*transition of polypeptide structures present in the

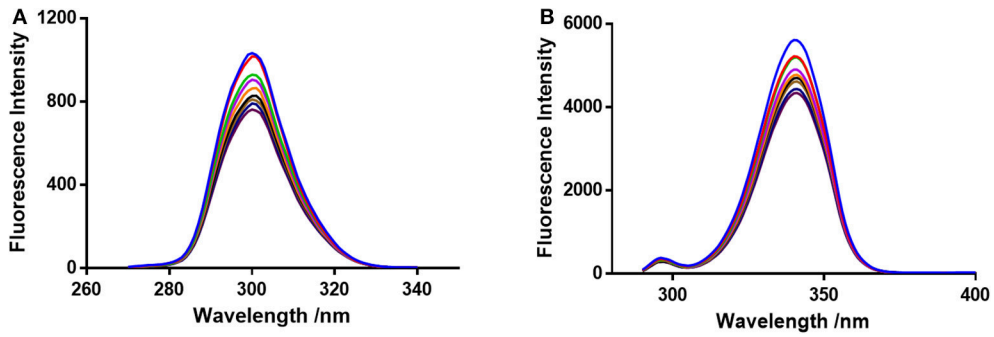



**FIGURE 4 |** Synchronous fluorescence spectroscopy of BSA and neratinib at 298 K **(A)** $\Delta\lambda = 15$ nm and **(B)** $\Delta\lambda = 60$ nm.

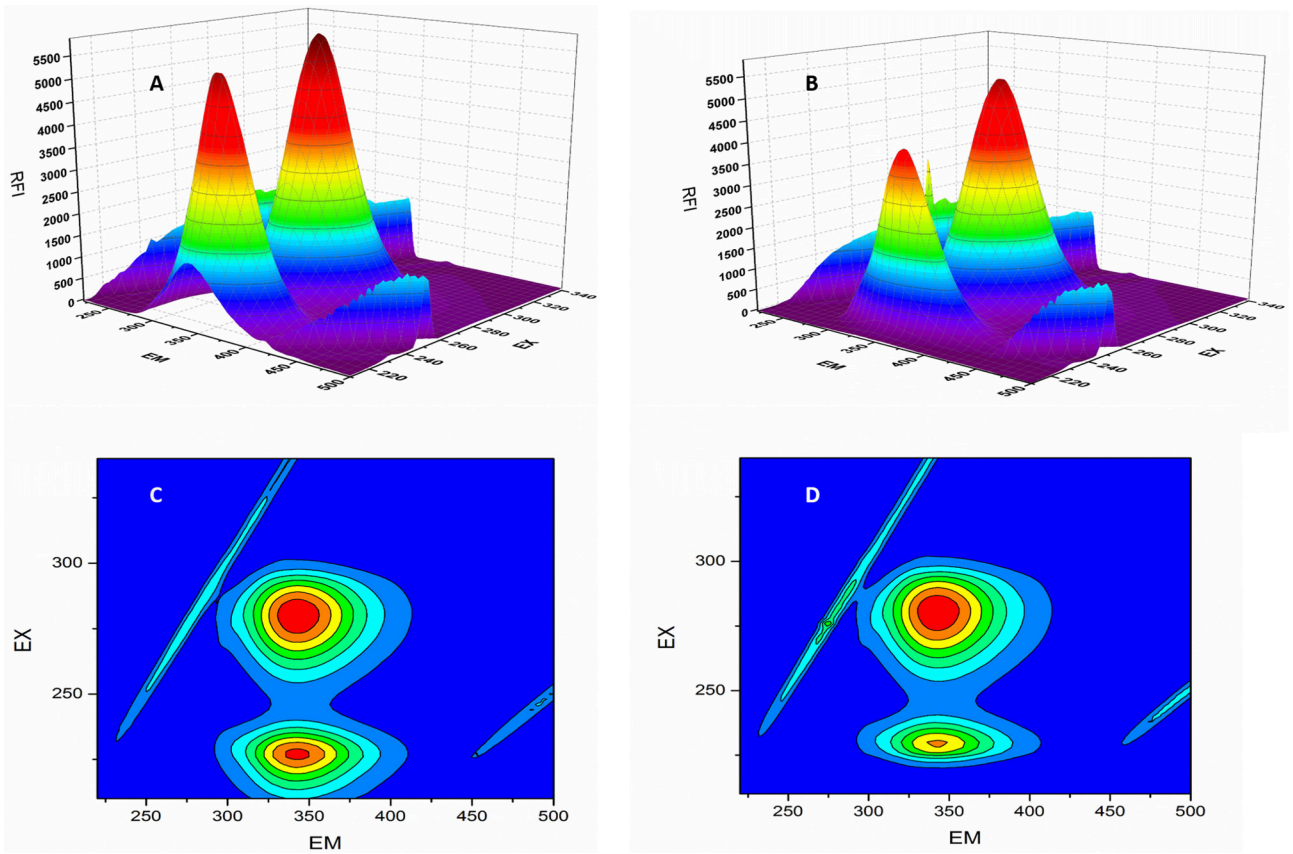



**FIGURE 5 |** 3D-spectroflurometric analysis of BSA and neratinib–BSA system. **(A,B)** are normal 3D spectra and **(C,D)** represent the contour plot of FI.

BSA molecule. Peak 2 was found at the excitation and emission wavelength of 280 and 342 nm, respectively. Tryptophan and tyrosine residues are responsible for the formation of Peak 2. A sharp decrease in the FI of BSA was witnessed after addition of neratinib meaning that fluorescence quenching occured. A sparse spectrum in the contour plot (**Figures 5C,D**) was observed for BSA in presence of neratinib, which confirms the occurrence of conformational changes in BSA after neratinib addition.

A decrease in the binding constants was noticed as the temperature increased indicating the instability of BSA-neratinib complex. Furthermore, the number of binding sites was found to be equal to 1, indicating a single class of binding sites on BSA.

Site specific probes, phenylbutazone and ibuprofen, were used for determination of the binding sites present on BSA (Hu et al., 2004). A decrease in the values of binding constants was observed in presence of drug site probes. Phenylbutazone caused a greater reduction in the binding constant compared

**TABLE 2 |** Binding and thermodynamic parameters of binding between neratinib and BSA.

| T(K) | R | Log $K_b \pm SD$ | $K_b \pm SD \times 10^4$ (L $mol^{-1}$) | n | ΔG (kJ $mol^{-1}$) | ΔH (kJ $mol^{-1}$) | ΔS ($Jmol^{-1}\cdot K^{-1}$) |
|---|---|---|---|---|---|---|---|
| 298 | 0.9938 | 4.909 | 8.10 ± 0.20 | 1.02 | −27.93 | −76.9 | −164 |
| 303 | 0.9905 | 4.653 | 4.50 ± 0.24 | 0.96 | −27.11 | | |
| 308 | 0.9902 | 4.383 | 2.42 ± 0.11 | 0.92 | −25.96 | | |

to ibuprofen inferring Site I as the binding site for neratinib (**Figure 3**).

## Types of Interaction Force Between BSA With Neratinib

The complex formation relies on the thermodynamic process due to the fact that binding constants are temperature-dependent. The thermodynamic processes help characterize the kind of forces engaged among BSA and neratinib (Ni et al., 2008). The forces that might be involved in binding small ligands to proteins include hydrogen bonds and Van der Waals forces, hydrophobic interaction or electrostatic forces. The binding mode is established based on the thermodynamic parameters that include enthalpy change ($\Delta H^0$), entropy change ($\Delta S^0$), and free energy change ($\Delta G^0$). The thermodynamic parameters were evaluated by the following equations:

$$\ln Kb = -\frac{\Delta H^0}{RT} + \frac{\Delta S^0}{R}$$

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 = -RT\ln Kb$$

$K_b$ and $R$ represent the binding constant and universal gas constant, respectively. The negative (–) $\Delta H^0$ and $\Delta S^0$ indicate the presence of hydrogen bonding and Van der Waals forces between BSA and neratinib. Moreover, $(-\Delta H^0)$ cannot occur during electrostatic interactions since these interactions occur when $\Delta H^0$ is either very small or almost zero (Ross and Subramanian, 1981; Ni et al., 2008). **Figure 3**, represents the van't Hoff plot for neratinib and BSA interaction. The spontaneous interaction between BSA and neratinib is indicated by $(-\Delta G^0)$ value. Both the enthalpy change and entropy change acquired negative values in the neratinib-BSA interaction, suggesting an enthalpy-driven interaction and the entropy value reported as negative number indicates its unfavorability for the binding process.

## UV–Vis Absorption Studies

The UV–vis absorption spectra suggests a complex formation occurred between BSA and neratinib (**Figure 6**). An increase in the absorption intensity of BSA was observed with higher neratinib concentrations. The complex formation between BSA and neratinib is further confirmed as a blue shift was observed in the λmax of BSA (Kandagal et al., 2006; Peng et al., 2015).

## FT-IR Studies

Infrared spectroscopy is used to investigate the secondary structures and dynamics of protein. The band frequencies as a result of amide I, II, and III vibrations in the IR region provide information about the secondary protein structure (i.e., the amide I band 1,600–1,700 $cm^{-1}$ and amide II band 1,548 $cm^{-1}$). The information provided by amide I is more valuable due to its sensitivity to protein structure change than amide II. **Figure 7** provides information regarding the changes in BSA after neratinib addition. It is clear that there were a shift of peak occurred in amide I from 1645.51 to 1652.88 $cm^{-1}$ and a slight shift in amide II peak from 1544.70 to 1543.02 $cm^{-1}$, suggesting a change in the secondary structure of BSA after interaction with neratinib.

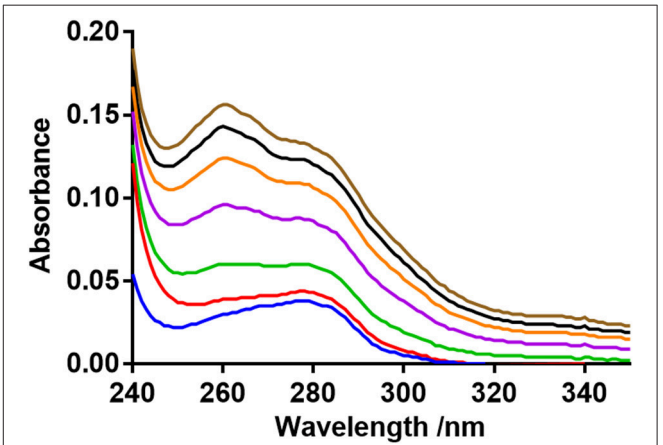



**FIGURE 6 |** BSA UV-absorption spectra in presence of neratinib.

## Molecular Simulation Studies

Molecular docking experiments were performed to understand the interaction between neratinib and BSA. The docking experiments further supported spectrophotometric and spectrofluorometric data (Ali et al., 2010; Shahabadi and Fili, 2014). In molecular docking studies, the ligand gets tied to the binding pocket of the protein in different positions thus providing valuable information on the binding site and mode. The two binding sites present on BSA protein are designated as Site I and Site II, and are present in sub-domains IIA and IIIA, respectively. The site probe experiment revealed site I as the binding site for neratinib which was further confirmed by the docking results. The sub-domains IIA of site 1 was analyzed with varied conformational adaptations and the least possible BSA-neratinib complex energies were obtained. **Figure 8A** represents the finest conformation of neratinib-BSA complex. It is evident that neratinib interacted with Trp-213 through pi-pi interaction

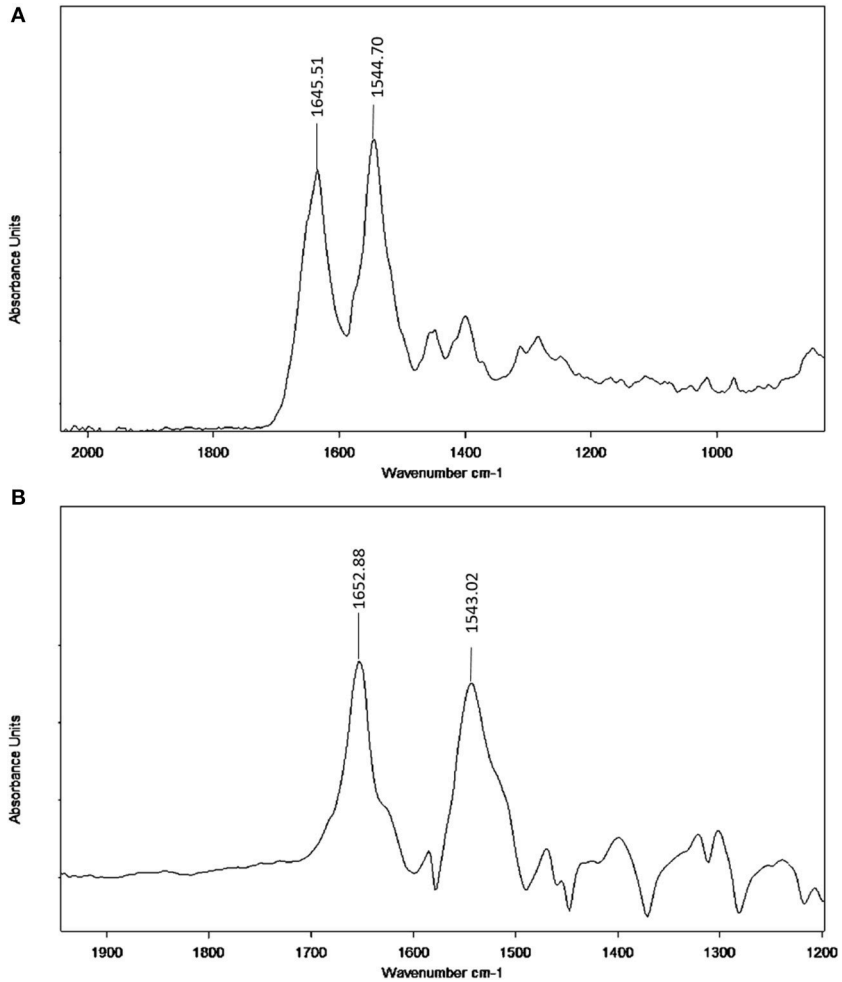



**FIGURE 7 |** FT-IR spectra **(A)** Free BSA in aqueous solution; **(B)** Difference spectra obtained by subtracting the spectrum of the neratinib-free form from that of the neratinib-bound form.

and with Asp-450 and Ala-209 by hydrogen bonds (**Figure 8B**). It was reported that neratinib forms a reversible covalent bond with Lys-190 of HSA. Neratinib contains a 4-(dimethylamino) crotonamide Michael acceptor and a covalent bond is formed between ε-amine of lysine of HSA and β-carbon of the amide functional group of neratinib. The covalent bond formed between neratinib and HSA is dependent on temperature, pH and time, and is independent of neratinib concentration (Chandrasekaran et al., 2010; Wang et al., 2010). The peptide LDELRDEGKASSAK is unique to human and monkey albumin; and neratinib binds to this peptide covalently. It has also been reported that neratinib does not bind covalently to plasma proteins from other species like dogs, rabbits and rodents as the sequence of amino acid residues from 182 to 195 in the albumin of these species is different than that in monkey and humans. The amino acid sequence of residues in BSA from 182 to 195 is ETMREKVLTSSARQ, meaning that BSA cannot bind covalently to neratinib due to this variation (Wang et al., 2010). The binding energy of neratinib-BSA complex at Site I by molecular docking was found to be $-24.12$ kj mol$^{-1}$, which is in an agreement with the binding energy of $-27.93$ kj mol$^{-1}$ found experimentally at 298 K. On the basis of experimental and docking results, it is concluded that hydrophobic (pi-pi interaction) and hydrophilic (hydrogen bonding) were involved in the BSA-neratinib complex stabilization.

## CONCLUSION

Neratinib approved for use in early stage HER2-overexpressed/amplified breast cancer was investigated for its interaction with BSA. The site probe and molecular docking experimental results established that neratinib binds to the site I, subdomain IIA of BSA. The fluorescence quenching, synchronous fluorescence, UV and FT-IR data together with the docking studies confirmed the formation of a complex between BSA and neratinib. Van der Waals forces and hydrogen bonding were found to be involved in the BSA-neratinib interaction in a enthalpy-driven manner. Based on

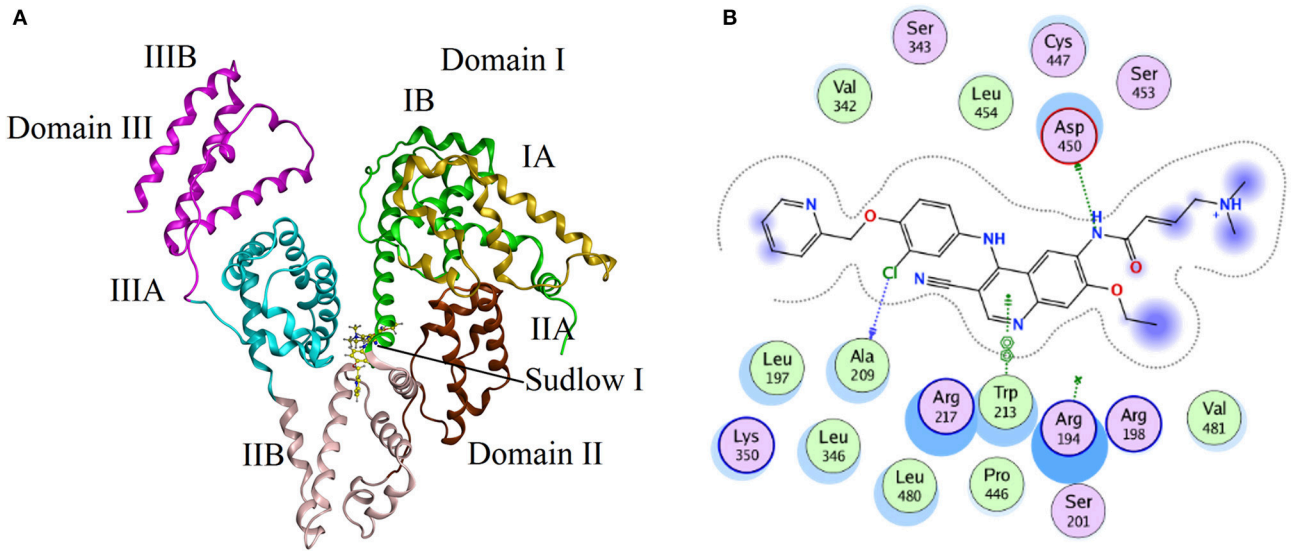



**FIGURE 8 | (A)** The docking conformation of neratinib-BSA complex with lowest energy; **(B)** Represents the amino acid residues that surround neratinib.

our findings, the pharmacological and biochemical aspects involved in the BSA-neratinib interaction could be better understood.

## AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: TW and MA. Performed the experiments: AB and SZ. Analyzed the data: AB and TW. Contributed reagents, materials, analysis tools: SZ, MA, and TW. Wrote the paper: TW and SZ.

## ACKNOWLEDGMENTS

The authors would like to extend their sincere appreciation to the Deanship of Scientific Research, King Saud University, for funding the research group No. RG-1438-042.

## REFERENCES

Albert, D. H., Tapang, P., Magoc, T. J., Pease, L. J., Reuter, D. R., Wei, R. Q., et al. (2006). Preclinical activity of ABT-869, a multitargeted receptor tyrosine kinase inhibitor. *Mol. Cancer Ther.* 5, 995–1006. doi: 10.1158/1535-7163.MCT-05-0410

Ali, H. I., Fujita, T., Akaho, E., and Nagamatsu, T. (2010). A comparative study of AutoDock and PMF scoring performances, and SAR of 2-substituted pyrazolotriazolopyrimidines and 4-substituted pyrazolopyrimidines as potent xanthine oxidase inhibitors. *J. Comput. Aided Mol. Des.* 24, 57–75. doi: 10.1007/s10822-009-9314-z

Berezhkovskiy, L. M. (2007). On the calculation of the concentration dependence of drug binding to plasma proteins with multiple binding sites of different affinities: determination of the possible variation of the unbound drug fraction and calculation of the number of binding sites of the protein. *J. Pharm. Sci.* 96, 249–257. doi: 10.1002/jps.20777

Bose, P., and Ozer, H. (2009). Neratinib: an oral, irreversible dual EGFR/HER2 inhibitor for breast and non-small cell lung cancer. *Expert Opin. Investig. Drugs* 18, 1735–1751. doi: 10.1517/13543780903305428

Burstein, H. J., Sun, Y., Dirix, L. Y., Jiang, Z., Paridaens, R., Tan, A. R., et al. (2010). Neratinib, an irreversible ErbB receptor tyrosine kinase inhibitor, in patients with advanced ErbB2-positive breast cancer. *J. Clin. Oncol.* 28, 1301–1307. doi: 10.1200/JCO.2009.25.8707

Chamani, J., and Heshmati, M. (2008). Mechanism for stabilization of the molten globule state of papain by sodium n-alkyl sulfates: spectroscopic and calorimetric approaches. *J. Colloid Interface Sci.* 322, 119–127. doi: 10.1016/j.jcis.2008.03.001

Chandrasekaran, A., Shen, L., Lockhead, S., Oganesian, A., Wang, J., and Scatina, J. (2010). Reversible covalent binding of neratinib to human serum albumin *in vitro*. *Drug Metab. Lett.* 4, 220–227. doi: 10.2174/187231210792928206

Chi, Z., Liu, R., Teng, Y., Fang, X., and Gao, C. (2010). Binding of oxytetracycline to bovine serum albumin: spectroscopic and molecular modeling investigations. *J. Agric. Food Chem.* 58, 10262–10269. doi: 10.1021/jf101417w

Feldinger, K., and Kong, A. (2015). Profile of neratinib and its potential in the treatment of breast cancer. *Breast Cancer.* 7, 147–162. doi: 10.2147/BCTT.S54414.

He, L. L., Wang, X., Liu, B., Wang, J., and Sun, Y. G. (2010). Interaction between ranitidine hydrochloride and bovine serum albumin in aqueous solution. *J. Solution Chem.* 39, 654–664. doi: 10.1007/s10953-010-9537-6

He, X. M., and Carter, D. C. (1992). Atomic structure and chemistry of human serum albumin. *Nature* 358, 209–215. doi: 10.1038/358209a0

Hu, Y. J., Liu, Y., Wang, J. B., Xiao, X. H., and Qu, S. S. (2004). Study of the interaction between monoammonium glycyrrhizinate and bovine serum albumin. *J. Pharm. Biomed. Anal.* 36, 915–919. doi: 10.1016/j.jpba.2004.08.021

Iqbal, N., and Iqbal, N. (2014). Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. *Mol. Biol. Int.* 2014:852748. doi: 10.1155/2014/852748

Jahanban-Esfahlan, A., Panahi-Azar, V., and Sajedi, S. (2015). Spectroscopic and molecular docking studies on the interaction between N-acetyl cysteine and bovine serum albumin. *Biopolymers* 103, 638–645. doi: 10.1002/bip.22697

Kandagal, P. B., Ashoka, S., Seetharamappa, J., Shaikh, S. M., Jadegoud, Y., and Ijare, O. B. (2006). Study of the interaction of an anticancer drug with human and bovine serum albumin: spectroscopic approach. *J. Pharm. Biomed. Anal.* 41, 393–399. doi: 10.1016/j.jpba.2005.11.037

Khorsand Ahmadi, S., Mahmoodian Moghadam, M., Mokaberi, P., Reza Saberi, M., and Chamani, J. (2015). A comparison study of the interaction between β-lactoglobulin and retinol at two different conditions: spectroscopic and molecular modeling approaches. *J. Biomol. Struct. Dyn.* 33, 1880–1898. doi: 10.1080/07391102.2014.977351

Kourie, H. R., Chaix, M., Gombos, A., Aftimos, P., and Awada, A., (2016). Pharmacodynamics, pharmacokinetics and clinical efficacy of neratinib in HER2-positive breast cancer and breast cancer with HER2 mutations. *Expert Opin. Drug Metab. Toxicol.* 12, 947–957. doi: 10.1080/17425255.2016.1198317

Kragh-Hansen, U. (1981). Molecular aspects of ligand binding to serum albumin. *Pharmacol. Rev.* 33, 17–53.

Lakowicz, J. R., and Weber, G. (1973). Quenching of fluorescence by oxygen. Probe for structural fluctuations in macromolecules. *Biochemistry* 12, 4161–4170. doi: 10.1021/bi00745a020

Marouzi, S., Sharifi Rad, A., Beigoli, S., Teimoori Baghaee, P., Assaran Darban, R. and Chamani, J. (2017). Study on effect of lomefloxacin on human holo-transferrin in the presence of essential and nonessential amino acids: spectroscopic and molecular modeling approaches. *Int. J. Biol. Macromol.* 97, 688–699. doi: 10.1016/j.ijbiomac.2017.01.047

Meti, M. D., Nandibewoor, S. T., Joshi, S. D., More, U. A., and Chimatadar, S. A. (2015). Multi-spectroscopic investigation of the binding interaction of fosfomycin with bovine serum albumin. *J. Pharm. Anal.* 5, 249–255. doi: 10.1016/j.jpha.2015.01.004

Ni, Y., Liu, G., and Kokot, S. (2008). Fluorescence spectrometric study on the interactions of Isoprocarb and sodium 2-isopropylphenate with bovine serum albumin. *Talanta* 76, 513–521. doi: 10.1016/j.talanta.2008.03.037

Peng, X., Wang, X., Qi, W., Huang, R., Su, R., and He, Z. (2015). Deciphering the binding patterns and conformation changes upon the bovine serum albumin–rosmarinic acid complex. *Food Funct.* 6, 2712–2726. doi: 10.1039/C5FO00597C

Ross, P. D., and Subramanian, S. (1981). Thermodynamics of protein association reactions: forces contributing to stability. *Biochemistry* 20, 3096–3102. doi: 10.1021/bi00514a017

Shahabadi, N., and Fili, S. M. (2014). Molecular modeling and multispectroscopic studies of the interaction of mesalamine with bovine serum albumin. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 118, 422–429. doi: 10.1016/j.saa.2013.08.110

Shen, G. F., Liu, T. T., Wang, Q., Jiang, M., and Shi, J. H. (2015). Spectroscopic and molecular docking studies of binding interaction of gefitinib, lapatinib and sunitinib with bovine serum albumin (BSA). *J. Photochem. Photobiol. B Biol.* 153, 380–390. doi: 10.1016/j.jphotobiol.2015.10.023

Shi, J. H., Wang, J., Zhu, Y. Y., and Chen, J. (2014). Characterization of interaction between isoliquiritigenin and bovine serum albumin: spectroscopic and molecular docking methods. *J. Lumin.* 145, 643–650. doi: 10.1016/j.jlumin.2013.08.042

US Food and Drug Administration (2018). *Drug Approval Database*. Available online at: https://www.fda.gov/drugs/informationondrugs/approveddrugs/ucm567259.htm.

Wang, J., Li-Chan, X. X., Atherton, J., Deng, L., Espina, R., Yu, L. et al. (2010). Characterization of HKI-272 covalent binding to human serum. *Drug Metab. Dispos.* 38, 1083–1093. doi: 10.1124/dmd.110.032292

Wani, T. A., AlRabiah, H., Bakheit, A. H., Kalam, M. A., and Zargar, S. (2017a). Study of binding interaction of rivaroxaban with bovine serum albumin using multi-spectroscopic and molecular docking approach. *Chem. Cent. J.* 11:134 doi: 10.1186/s13065-017-0366-1

Wani, T. A., Bakheit, A. H., Al-Majed, A. R. A., Bhat, M. A., and Zargar, S. (2017b). Study of the interactions of bovine serum albumin with the new anti-inflammatory agent 4-(1, 3-Dioxo-1, 3-dihydro-2H-isoindol-2-yl)-N′-[(4-ethoxy-phenyl) methylidene] benzohydrazide using a multi-spectroscopic approach and molecular docking. *Molecules* 22:1258. doi: 10.3390/molecules22081258

Wani, T. A., Bakheit, A. H., Zargar, S., Hamidaddin, M. A., and Darwish, I. A. (2017c). Spectrophotometric and molecular modelling studies on *in vitro* interaction of tyrosine kinase inhibitor linifanib with bovine serum albumin. *PLoS ONE* 12:e0176015. doi: 10.1371/journal.pone.0176015

Wani, T. A., Zargar, S., and Ahmad, A. (2015). Ultra performance liquid chromatography tandem mass spectrometric method development and validation for determination of neratinib in human plasma. *S. Afr. J. Chem.* 68, 93–98. doi: 10.17159/0379-4350/2015/v68a14

Xiao, J., Wu, M., Kai, G., Wang, F., Cao, H., and Yu, X. (2011). ZnO-ZnS QDs interfacial heterostructure for drug and food delivery application: enhancement of the binding affinities of flavonoid aglycones to bovine serum albumin. *Nanomedicine* 7, 850–858. doi: 10.1016/j.nano.2011.02.003

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 07 May 2018
doi: 10.3389/fchem.2018.00154

# Discovery of the Linear Region of Near Infrared Diffuse Reflectance Spectra Using the Kubelka-Munk Theory

*Shengyun Dai, Xiaoning Pan, Lijuan Ma, Xingguo Huang, Chenzhao Du, Yanjiang Qiao* and Zhisheng Wu**

*Key Laboratory of TCM-Information Engineering of State Administration of TCM, Pharmaceutical Engineering and New Drug Development of Traditional Chinese, Medicine of Ministry of Education, Beijing University of Chinese Medicine, Beijing, China*

OPEN ACCESS

**Edited by:**
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

**Reviewed by:**
*Eleonora-Mihaela Ungureanu,*
*Politehnica University of Bucharest, Romania*
*Michalina Kotyczka-Moranska,*
*Institute for Chemical Processing of Coal, Poland*

***Correspondence:**
*Yanjiang Qiao*
*yjqiao@263.net*
*Zhisheng Wu*
*wzs@bucm.edu.cn*

**Specialty section:**
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

**Received:** *30 November 2017*
**Accepted:** *19 April 2018*
**Published:** *07 May 2018*

**Citation:**
*Dai S, Pan X, Ma L, Huang X, Du C, Qiao Y and Wu Z (2018) Discovery of the Linear Region of Near Infrared Diffuse Reflectance Spectra Using the Kubelka-Munk Theory. Front. Chem. 6:154. doi: 10.3389/fchem.2018.00154*

Particle size is of great importance for the quantitative model of the NIR diffuse reflectance. In this paper, the effect of sample particle size on the measurement of harpagoside in *Radix Scrophulariae* powder by near infrared diffuse (NIR) reflectance spectroscopy was explored. High-performance liquid chromatography (HPLC) was employed as a reference method to construct the quantitative particle size model. Several spectral preprocessing methods were compared, and particle size models obtained by different preprocessing methods for establishing the partial least-squares (PLS) models of harpagoside. Data showed that the particle size distribution of 125–150 μm for *Radix Scrophulariae* exhibited the best prediction ability with $R^2_{pre} = 0.9513$, RMSEP = 0.1029 mg·g$^{-1}$, and RPD = 4.78. For the hybrid granularity calibration model, the particle size distribution of 90–180 μm exhibited the best prediction ability with $R^2_{pre} = 0.8919$, RMSEP = 0.1632 mg·g$^{-1}$, and RPD = 3.09. Furthermore, the Kubelka-Munk theory was used to relate the absorption coefficient *k* (concentration-dependent) and scatter coefficient *s* (particle size-dependent). The scatter coefficient *s* was calculated based on the Kubelka-Munk theory to study the changes of *s* after being mathematically preprocessed. A linear relationship was observed between *k*/*s* and absorption *A* within a certain range and the value for *k*/*s* was >4. According to this relationship, the model was more accurately constructed with the particle size distribution of 90–180 μm when *s* was kept constant or in a small linear region. This region provided a good reference for the linear modeling of diffuse reflectance spectroscopy. To establish a diffuse reflectance NIR model, further accurate assessment should be obtained in advance for a precise linear model.

**Keywords: Kubelka-Munk theory, Near infrared (NIR) diffuse reflectance spectroscopy, particle size, PLS, harpagoside, *Radix Scrophulariae***

## INTRODUCTION

The implementation of process analytical technology (PAT) in the pharmaceutical industry is intended to enhance the quality of products through the measurement of critical quality and performance parameters (Roggo and Ulmschneider, 2008). Near infrared spectroscopy (NIRS) is regarded as a vital tool for the implementation of PAT, as it is increasingly used in pharmaceutical research and development due to its high analysis speed, low-cost, and non-destructive characteristics (De Beer et al., 2011). NIR spectra of chemical species (consisting of C–H, N–H, O–H, and S–H bonds; Sarraguça et al., 2011) can be used to predict their chemical and physical properties (Prieto et al., 2009).

The NIR technology includes two main parts that are transmission spectroscopy and diffuse reflectance spectroscopy. The selection of spectral form is mainly based on the state of samples (i.e., transmission spectroscopy is suitable for liquid samples such as herbal extracts and liquid preparations, while diffuse reflectance spectroscopy is generally used for solid samples such as pharmaceutical powders or granules). Diffuse reflectance spectroscopy is an analytical technique that measures the diffuse reflection of different wavelengths of light to obtain the surface information of the materials.

Various physical, chemical, and biochemical properties in Mediterranean soils were NIR predicted (Zornoza et al., 2008). Chen et al. employed an NIR model for the analysis of total polyphenol content in green tea (Chen Q. et al., 2008). Classification accuracy of about 100 % was obtained by discriminant and classification tree analyses of 82 honey samples by diffuse reflectance mid-infrared Fourier transform spectroscopy (DRIFTS) (Bertelli et al., 2007). Borin et al. utilized NIR technology for the simultaneous quantification of some common adulterants (starch, whey, or sucrose) found in milk powder samples (Borin et al., 2006). All these investigations have illustrated the trend of using NIR technology to predict physical and chemical information.

Recently, the application of NIR in studying Chinese herbal medicine (CHM) has dramatically increased such as discrimination analysis and quality control for various samples e.g., raw materials, excipients, and dosage forms. Wu et al. used the NIR and different PLS models to quantify the baicalin contents of Yinhuang oral solution based on a total error concept (Wu et al., 2013). Chen et al. employed NIR to distinguish Ganoderma lucidum samples collected from different geographical origins using principal component analysis (PCA) and discriminant analysis algorithms (Chen Y. et al., 2008).

**Abbreviations:** NIR, Near Infrared Diffuse; PAT, Process Analytical Technology; DRIFTS, Diffuse Reflectance Mid-infrared Fourier Transform Spectroscopy; CHM, Chinese Herbal Medicine; NIRS, NIR Spectroscopy; MVA, Multivariate Data Analysis; HPLC, High Performance Liquid Chromatography; RMSEC, Root Mean Square Error of Calibration; RMSECV; Root Mean Square Error of Cross-Validation; RMSEP, Root Mean Square Error of Prediction; MSC, Multiplicative Scatter Correction; SNV, Standard Normal Variate; 1D, First Derivative; 2D, Second Derivative; SG, Savitzky–Golay; PRESS, Predicted Residual Sum of Squares; PLS, Partial Least Squares; SCOT, Second Overtones Region; FCOT, First Combination-Overtone; RPD, Residual Predictive Deviation; API, Active Pharmaceutical Ingredient; EMSC, Extended Multiplicative Scatter Correction.

On the other hand, it is well known that the particle size of sample affects NIR spectra. Several studies have been published on the effect of particle size on the determination of drug content in mixed powder products (Norris and Williams, 1984; Aucott and Garthwaite, 1988; Bull, 1991). Franke et al. (1998) reported the particle size determination of lactose using chemometrics-based NIR spectra. However, they did not mention any basic principle to determine particle size in the experiments. Paskatan et al. (2001) reviewed theoretical and practical particle size analysis of powder by NIR spectroscopy. But they did not show the relationship between the basic light scattering principle and the particle size of main contents.

Kubelka-Munk theory (Otsuka, 2004) is the basic quantitative theory of NIRS. The particle size of sample affects the light scattering, directly influencing model construction. It was shown that an accurate knowledge of the particles is crucial in the product development (Blanco and Peguero, 2008). Meanwhile, the differences in CHM particle size could result in different optical path lengths and multiplicative light scattering effects (Jin et al., 2012). Thus, it is important to establish an expeditious method to determine the particle size of CHM.

However, there were a few NIR studies on the simultaneous determination of particle size and active pharmaceutical ingredients of CHM. Wu Z. S. et al. demonstrated that the particle size affected NIR measurement of saikosaponin A in *Bupleurum chinense DC* (Wu et al., 2015). Bittner et al. employed a successful application of NIR spectroscopy in combination with multivariate data analysis (MVA) for the simultaneous identification and particle size determination of amoxicillin trihydrate particles (Bittner et al., 2011).

*Scrophularia* radix (Xuanshen), the root of *Scrophularia ningpoensis Hemsl.*, was a typical CHM with a history going back over 1000 years (The State Pharmacopoeia Commission of People's Republic of China, 2015). It is originally from Zhejiang province and it is a component of the natural herbal supplement named "Zhe Ba Wei." The major ingredients of *Scrophularia* radix are iridoids, and harpagoside is one of the main bioactive components with antioxidant, antimicrobial and antitumor activities (Miyazawa and Okuno, 2003; Jing et al., 2011).

In this study, *Scrophularia* radix was taken as an example and harpagoside was regarded as an API of *Scrophularia* radix. HPLC was used as a reference method to determine the harpagoside

**TABLE 1 |** HPLC gradient elution of *Scrophularia* radix extract.

| Time/min | A/% | B/% |
|---|---|---|
| 0–10 | 5–10 | 95-90 |
| 10–25 | 10–33 | 90–67 |
| 25–35 | 33–50 | 67–50 |
| 35–40 | 50–60 | 50–40 |
| 40–45 | 60–70 | 40–30 |
| 45–55 | 70–80 | 30–20 |
| 55–60 | 80–5 | 20–95 |

content. NIR was used to monitor the prediction potential of the models of single particle size and mix particle size simultaneously. To our best knowledge, this paper is the first to study on particle size and harpagoside determination in *Scrophularia* radix with NIR diffuse reflectance spectroscopy. The differences between single particle size model and mix particle size model from the perspective of the Kubelka-Munk theory were explained.

# MATERIALS AND METHOD

## Materials

Ten batches of *S. ningpoensis* Hemsl. radix were gifted from Daozhen (Guizhou, China), three representative samples were taken from each batch. All samples were identified by Prof. Chunsheng Liu (Beijing University of Chinese Medicine, China). Harpagoside reference standard (lot: 111730-201307) was purchased from the National Institutes for Food and Drug Control (Beijing, China). Acetonitrile (Fisher Scientific, Pittsburgh, PA) was of HPLC-grade. Acetic acid (Beijing Chemical Works, Beijing, China) was of analytical grade. Deionised water was purchased from Hangzhou Wahaha Co., Ltd (Zhejiang, China).

## Preparation of Samples

*Scrophularia* radix samples were crushed into pieces by a disintegrator after brushing off soil dust from the surface. Thirty samples of *Scrophularia* radix were then pulverized with a blender and screened through a 10-mesh sieve. Finally, the powders were divided into four parts. One part was used for HPLC determination of the harpagoside content. The remaining parts were then smashed and screened through 24-, 50-, 65-, 80-, 100-, 120-, and 150-mesh sieves.

An amount of each sieved sample of *Scrophularia* radix powder (1 g) was accurately weighed and placed in a 100-mL Erlenmeyer flask. The sample was extracted with 50 mL of 50% ethanol under ultrasonic vibration (40 kHZ, 220 V) for 45 min. After cooling to room temperature, the solution was filtered through a 0.45-μm membrane filter for HPLC analysis.

## NIR Equipment and Measurement

The NIR spectra were recorded by a XDS Rapid Content Analyser and VISION software (Metrohm NIR Systems, Florida, USA).

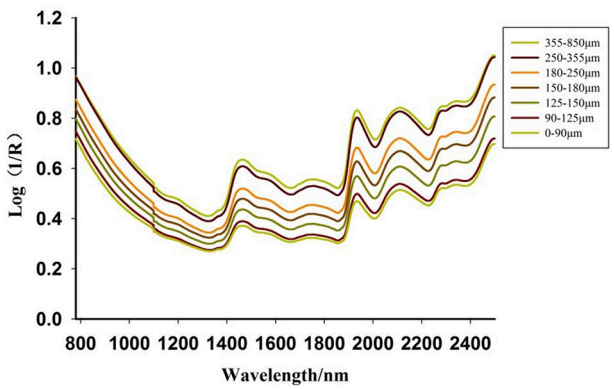



**FIGURE 1 |** NIR diffuse reflectance spectra of Scrophularia radix.

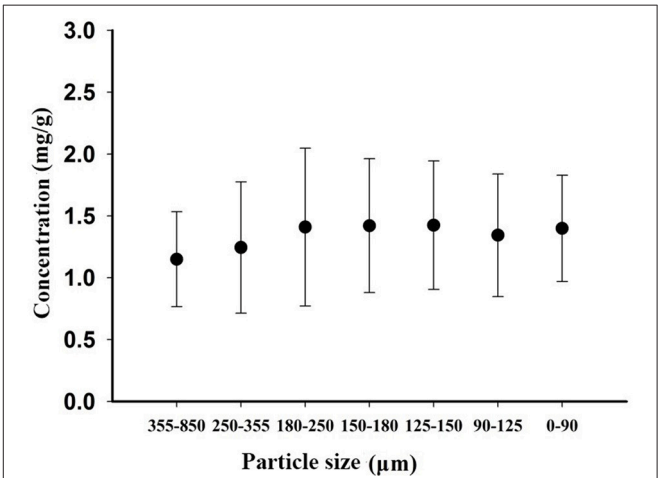



**FIGURE 3 |** Harpagoside concentration of 30 samples of different particle sizes.

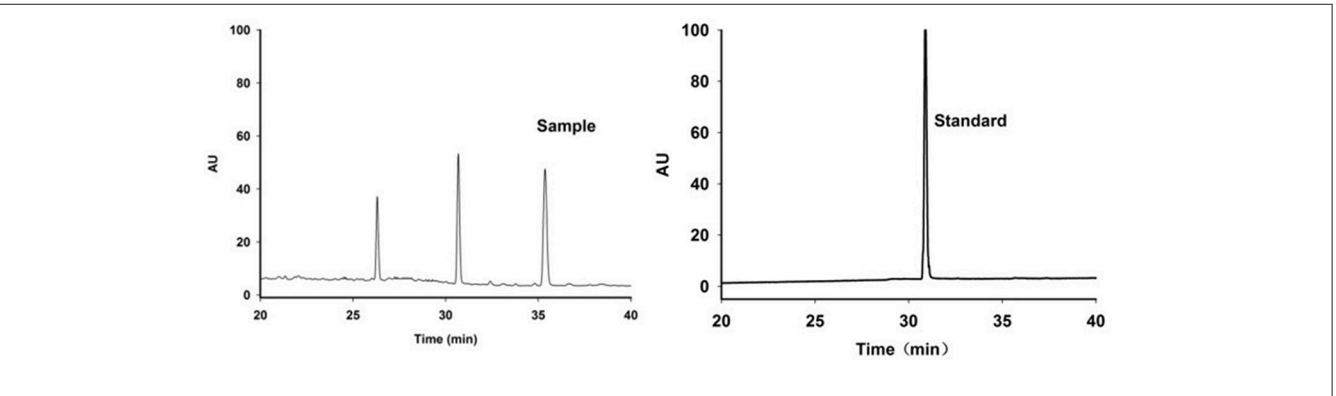



**FIGURE 2 |** Representative HPLC chromatograms of Scrophularia radix sample and harpagoside standard.

The wavelength range for the spectra was 780–2,500 nm. Each spectrum was an average of 64 scans with air as the background, and the wavelength increment was of 0.5 nm. Unless stated otherwise, each sample was measured in triplicate and its mean value was used in the subsequent analysis.

## HPLC Method

A certain amount of harpagoside standard was accurately weighed with an XS205DU electronic balance (Mettler Toledo, Greifensee, Switzerland) and then dissolved in 100 mL of methanol to obtain the concentration of 0.02432 mg·mL$^{-1}$.

HPLC analysis of *Scrophularia* radix (according to Chinese Pharmacopoeia, 2010 ed) was carried out using a Waters 2695 HPLC system, Waters 2996 DAD detector and auto-sampler (Waters Technologies, Palo Alto, CA). Ten microliters aliquots of the sample solutions were chromatographically analyzed in gradient elution mode on an octadecylsilyl column [250 × 4.6 mm, 5 μm (Dikma, China)] with the mobile phase consisting

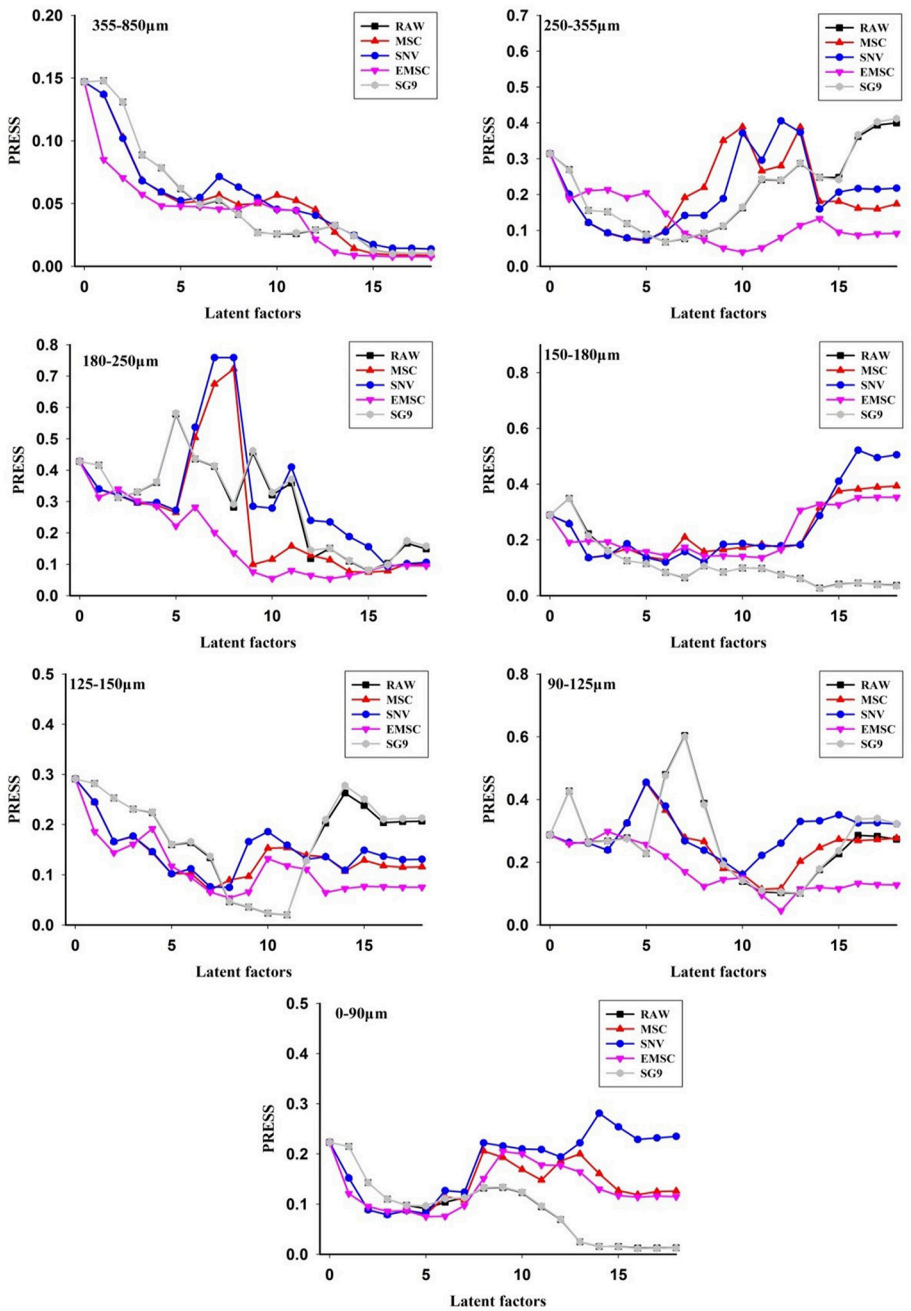



**FIGURE 4 |** The PRESS values of different preprocessing methods for single particle size model.

of acetonitrile and 0.4% acetic acid (*v*/*v*) at a flow rate of 1.0 mL·min$^{-1}$ (**Table 1**). The column temperature was kept at 30°C and the detection wavelength set at 280 nm. This chromatographic method exhibited good linearity (Y = 3 × 10$^6$X−104747, $R^2$ = 0.9998) over the concentration range 0.04864–0.02432 mg·mL$^{-1}$.

## Software

Data analysis was performed by the Unscrambler version 9.6 software package (CAMO Software AS, Oslo, Norway) and home-made routines programmed in MATLAB code (MATLAB v7.0, Math Works, Natick, MA). Following the Kennard-Stone algorithm, 210 samples were divided into 140 calibration samples and 70 validation samples. The root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV), root mean square error of prediction (RMSEP) and corresponding $R^2$ were used to evaluate the PLS model.

In order to establish a robust harpagoside model, a number of preprocessing methods were selected. For instance, multiplicative scatter correction (MSC) and standard normal variate (SNV) were used to eliminate redundant effects of

**TABLE 2 |** PLS model using preprocessing methods for different single particle sizes.

| Particle size (μm) | Preprocessing | Model evaluation parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | RMSEC | $R^2$ | RMSECV | $R^2$ | RMSEP | $R^2$ | RPD |
| 355–850 | RAW# | 0.0576 | 0.9750 | 0.1642 | 0.8167 | 0.2094 | 0.7279 | 2.02* |
| | MSC | 0.1451 | 0.8414 | 0.2248 | 0.6568 | 0.2187 | 0.7031 | 1.93 |
| | SNV | 0.1450 | 0.8418 | 0.2288 | 0.6444 | 0.2169 | 0.7082 | 1.95 |
| | EMSC | 0.1345 | 0.8638 | 0.2194 | 0.6730 | 0.2194 | 0.7014 | 1.93 |
| | SG9 | 0.0575 | 0.9751 | 0.1635 | 0.8184 | 0.2098 | 0.7269 | 2.01 |
| 250–355 | RAW | 0.1497 | 0.9208 | 0.2601 | 0.7843 | 0.1884 | 0.8541 | 2.76 |
| | MSC | 0.1701 | 0.8978 | 0.2657 | 0.7750 | 0.2050 | 0.8272 | 2.54 |
| | SNV | 0.1704 | 0.8974 | 0.2715 | 0.7651 | 0.2051 | 0.8270 | 2.53 |
| | EMSC | 0.0625 | 0.9862 | 0.1996 | 0.8730 | 0.1643 | 0.8890 | 3.16* |
| | SG9 | 0.1498 | 0.9208 | 0.2602 | 0.7841 | 0.1885 | 0.8540 | 2.76 |
| 180–250 | RAW | 0.1050 | 0.9714 | 0.5666 | 0.2497 | 0.1729 | 0.9265 | 3.89 |
| | MSC | 0.0839 | 0.9818 | 0.3406 | 0.7289 | 0.2840 | 0.8017 | 2.37 |
| | SNV | 0.0678 | 0.9881 | 0.5278 | 0.3489 | 0.4332 | 0.5387 | 1.55 |
| | EMSC | 0.0800 | 0.9834 | 0.2339 | 0.8721 | 0.2198 | 0.8812 | 3.06 |
| | SG9 | 0.1074 | 0.9701 | 0.5736 | 0.2312 | 0.1709 | 0.9281 | 3.93* |
| 150–180 | RAW | 0.0304 | 0.9965 | 0.3150 | 0.6562 | 0.1699 | 0.9038 | 3.40 |
| | MSC | 0.2911 | 0.6746 | 0.3686 | 0.5291 | 0.3783 | 0.5232 | 1.53 |
| | SNV | 0.1566 | 0.9058 | 0.3459 | 0.5854 | 0.2484 | 0.7945 | 2.33 |
| | EMSC | 0.1436 | 0.9208 | 0.3801 | 0.4992 | 0.2609 | 0.7733 | 2.21 |
| | SG9 | 0.0300 | 0.9965 | 0.3154 | 0.6553 | 0.1696 | 0.9041 | 3.40* |
| 125–150 | RAW | 0.0362 | 0.9950 | 0.1537 | 0.9189 | 0.2224 | 0.7726 | 2.21 |
| | MSC | 0.1172 | 0.9477 | 0.2630 | 0.7623 | 0.1470 | 0.9006 | 3.34 |
| | SNV | 0.1082 | 0.9554 | 0.2760 | 0.7384 | 0.1029 | 0.9513 | 4.78* |
| | EMSC | 0.0777 | 0.9770 | 0.2324 | 0.8145 | 0.1247 | 0.9285 | 3.94 |
| | SG9 | 0.0362 | 0.9950 | 0.1553 | 0.9171 | 0.2225 | 0.7722 | 2.21 |
| 90–125 | RAW | 0.0644 | 0.9840 | 0.3724 | 0.5164 | 0.1722 | 0.7574 | 2.14 |
| | MSC | 0.0604 | 0.9859 | 0.4020 | 0.4365 | 0.1460 | 0.8257 | 2.52 |
| | SNV | 0.0612 | 0.9855 | 0.4016 | 0.4376 | 0.1728 | 0.7557 | 2.13 |
| | EMSC | 0.0833 | 0.9732 | 0.3505 | 0.5718 | 0.1655 | 0.7760 | 2.23 |
| | SG9 | 0.0651 | 0.9836 | 0.3768 | 0.5049 | 0.1715 | 0.7596 | 2.15 |
| < 90 | RAW | 0.0620 | 0.9809 | 0.3505 | 0.4493 | 0.1298 | 0.8600 | 2.82* |
| | MSC | 0.2352 | 0.7252 | 0.2808 | 0.6466 | 0.3437 | 0.0175 | 1.06 |
| | SNV | 0.2352 | 0.7253 | 0.2810 | 0.6460 | 0.3444 | 0.0133 | 1.06 |
| | EMSC | 0.1560 | 0.8791 | 0.2745 | 0.6623 | 0.2471 | 0.4920 | 1.48 |
| | SG9 | 0.0627 | 0.9805 | 0.3523 | 0.4437 | 0.1302 | 0.8590 | 2.81 |

*#The original spectra without any pretreatment.*

**The best preprocessing methods using in each different single particle size.*

particle size. Derivative methods including first derivative (1D) and second derivative (2D) were obtained to reduce baseline variations observed in original diffuse reflectance spectra and to enhance spectral features. Meanwhile, a nine-point Savitzky-Golay smoothing filter (SG) was employed to depress the background noise amplified by the derivative. For the particle size model, MSC, SNV, and second derivative were not appropriate for an effect to be modeled, so 1D + SG, normalization and baseline subtraction were used. Leave-one-out cross-validation was used to validate the validity of methods. The lowest predicted residual sum of squares (PRESS) value was used to determine the optimum latent variables.

## Quantitative Models of NIR Diffuse Reflectance Using the Kubelka-Munk Theory

Kubelka-Munk theory is the theoretical basis for the establishment of quantitative models of NIR diffuse reflectance and its function is as follows (Otsuka, 2004):

$$f(R_\infty) = \frac{(1-R_\infty)^2}{2R_\infty} = \frac{k}{s}$$

According to the Kubelka-Munk function, reflectance is inversely to proportional to the light-scattering coefficient ($s$), and the $s$ value is inversely proportional to particle size.

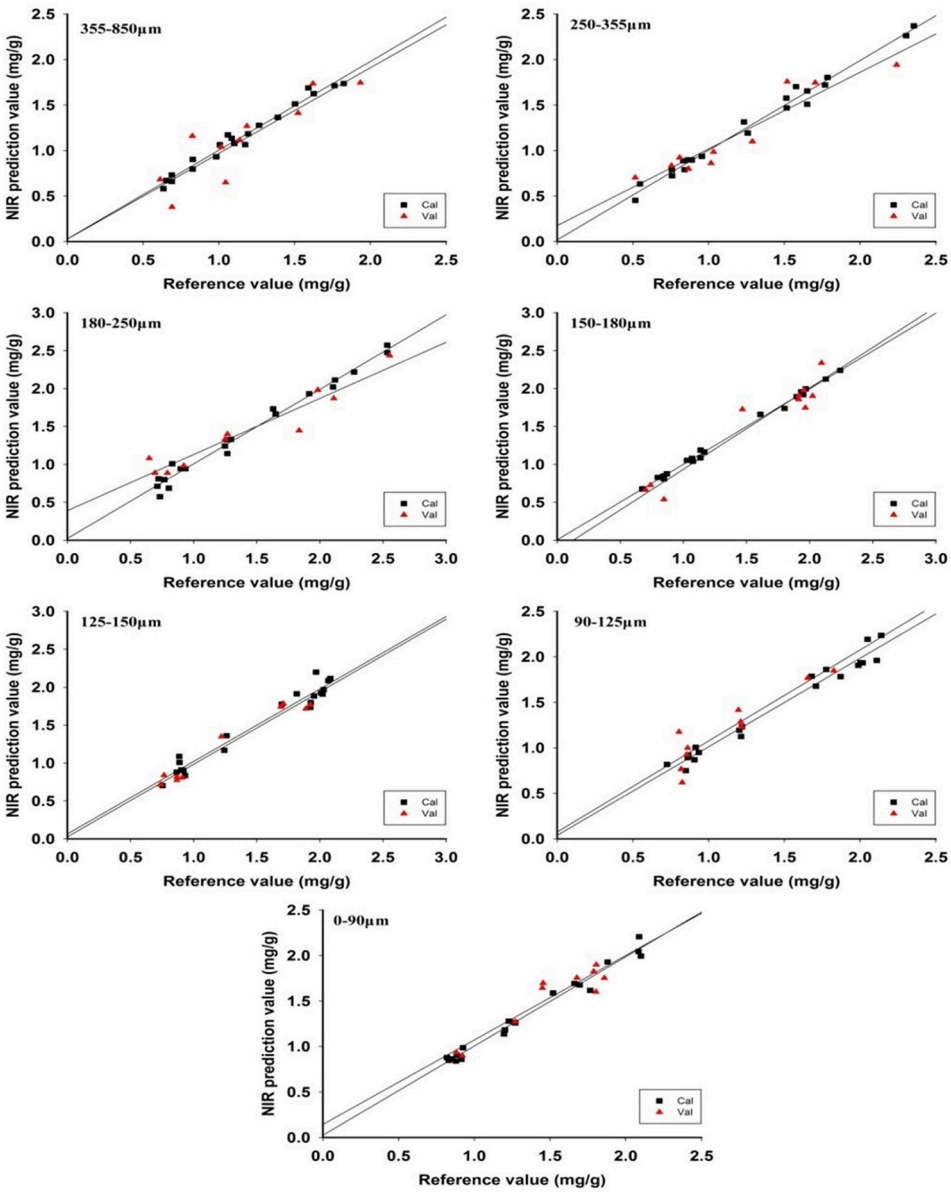



**FIGURE 5 |** The relation map of the reference value and predicted value using each different particle size.

The absorbance of NIR diffuse reflectance is expressed by the Kubelka-Munk equation:

$$A = -lg\left[1 + \frac{k}{s} - \sqrt{\left(\frac{k}{s}\right)^2 + 2\left(\frac{k}{s}\right)}\right]$$

# RESULTS AND DISCUSSION

## Spectral Characteristics of NIR Diffuse Reflectance Spectra of Different Particle Size Samples

The representative raw spectra of *Scrophularia* radix with different particle sizes are shown in **Figure 1** i.e., the spectral profiles were similar in shape. However, the main influences of particle size variation on diffuse reflectance spectra was the baseline offset. The well-known phenomenon that larger particles showed a stronger absorption, illustrates that the particle size is vital to the response. Some weak absorption peaks were demonstrated in the second overtone region (SCOT, 1,000–1,400 nm) of the fundamental C-H stretching bands, while much fluctuations in the region of first combination-overtone (FCOT, 1,400–2,040 $cm^{-1}$) and combination region (CR, 2,040-2,500 nm) were observed. Those absorption peaks might be caused by the diffuse reflectance on different particle sizes.

## HPLC Determination of Harpagoside Content in *Scrophularia* Radix

The HPLC chromatograms of the representative sample and standard are shown in **Figure 2**. The retention time of harpagoside in a sample extract was the same as that for the standard solution. **Figure 3** shows the harpagoside concentration of 30 samples. There is a significant difference in harpagoside concentration of samples of different particle sizes. The biggest difference of the particle sizes was located in the range of 180–250 μm, but the overall concentration design was suitable for the modeling.

**TABLE 3 |** Preprocessing methods for different mix particle size models (3 particle size ranges).

| Mix particle size (μm) | Preprocessing | Model evaluation parameters | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | RMSEC | $R^2$ | RMSECV | $R^2$ | RMSEP | $R^2$ | RPD |
| 180–850 | RAW# | 0.2492 | 0.7777 | 0.3188 | 0.6482 | 0.2699 | 0.7426 | 2.00* |
| | MSC | 0.2514 | 0.7737 | 0.3392 | 0.6016 | 0.3002 | 0.6817 | 1.80 |
| | SNV | 0.2861 | 0.7068 | 0.3273 | 0.6291 | 0.3172 | 0.6446 | 1.70 |
| | EMSC | 0.2319 | 0.8075 | 0.3410 | 0.5973 | 0.2781 | 0.7268 | 1.95 |
| | SG9 | 0.2494 | 0.7773 | 0.3193 | 0.6471 | 0.2702 | 0.7421 | 2.00 |
| 150–355 | RAW | 0.1927 | 0.8813 | 0.2338 | 0.8311 | 0.2157 | 0.8639 | 2.76 |
| | MSC | 0.2039 | 0.8671 | 0.2752 | 0.7659 | 0.2594 | 0.8033 | 2.29 |
| | SNV | 0.2654 | 0.7748 | 0.3081 | 0.7065 | 0.3117 | 0.7159 | 1.91 |
| | EMSC | 0.1467 | 0.9312 | 0.2348 | 0.8296 | 0.2053 | 0.8767 | 2.90* |
| | SG9 | 0.1933 | 0.8805 | 0.2345 | 0.8300 | 0.2161 | 0.8634 | 2.75 |
| 125–250 | RAW | 0.1592 | 0.9175 | 0.2233 | 0.8430 | 0.2592 | 0.7923 | 2.23 |
| | MSC | 0.1691 | 0.9069 | 0.2358 | 0.8250 | 0.2473 | 0.8109 | 2.34 |
| | SNV | 0.1684 | 0.9077 | 0.2428 | 0.8145 | 0.2902 | 0.7397 | 1.99 |
| | EMSC | 0.1646 | 0.9118 | 0.2358 | 0.8251 | 0.2408 | 0.8208 | 2.40* |
| | SG9 | 0.1597 | 0.9171 | 0.2239 | 0.8423 | 0.2595 | 0.7918 | 2.23 |
| 90–180 | RAW | 0.1395 | 0.9266 | 0.1983 | 0.8565 | 0.1843 | 0.8623 | 2.74 |
| | MSC | 0.1721 | 0.8881 | 0.2538 | 0.7649 | 0.1926 | 0.8495 | 2.62 |
| | SNV | 0.1805 | 0.8771 | 0.2706 | 0.7327 | 0.1978 | 0.8413 | 2.55 |
| | EMSC | 0.1572 | 0.9067 | 0.2342 | 0.7998 | 0.1632 | 0.8919 | 3.09* |
| | SG9 | 0.1393 | 0.9268 | 0.1974 | 0.8577 | 0.1844 | 0.8621 | 2.74 |
| 0–150 | RAW | 0.1585 | 0.8975 | 0.2009 | 0.8407 | 0.1744 | 0.8272 | 2.45 |
| | MSC | 0.1567 | 0.8998 | 0.2315 | 0.7886 | 0.1699 | 0.8359 | 2.51* |
| | SNV | 0.1655 | 0.8883 | 0.2499 | 0.7536 | 0.1757 | 0.8245 | 2.43 |
| | EMSC | 0.1553 | 0.9016 | 0.2268 | 0.7970 | 0.1736 | 0.8287 | 2.46 |
| | SG9 | 0.1588 | 0.8971 | 0.2008 | 0.8409 | 0.1746 | 0.8268 | 2.44 |

*#The original spectra without any pretreatment.*

**The best preprocessing method for different mix particle size models.*

## PLS Models for NIR Diffuse Reflectance Data Using *Scrophularia* Radix of Each Single Particle Size

Based on different preprocessing methods, the PLS model for each particle size was constructed. **Figure 4** showed the relationship between the latent variables and PRESS for different preprocessing methods. In general, the lowest PRESS value means the best latent variables (Pan et al., 2015). The model was validated for prediction by internal sample set. Moreover, the model performance values for each particle size using different preprocessing methods are illustrated in **Table 2**. Data showed that the raw spectra were the best to construct the particle size model of 355–850 μm and <90 μm. While the best preprocessing method for the particle size model of 250–355 μm, 180–250 μm, 150–180 μm, 125–150 μm, and 90-150 μm was EMSC, SG9, SG9, SNV, and MSC, respectively.

In addition, the model evaluation parameters, i.e., RMSEC, RMSECV, RMSEP, and RPD, for the particle size of 355–850 μm was 0.0576, 0.1642, 0.2094, and 2.02, respectively. The parameter values of other particle sizes are summarized in **Table 2**. The relation map between predicted value and reference value is shown in **Figure 5**, indicating that the best prediction result was for the particle size of 125–150 μm. Therefore, it could be known that the NIR model was influenced by different particle sizes and its quantitative characteristics was explored according to different particle sizes.

## PLS Models for NIR Diffuse Reflectance Data Using *Scrophularia* Radix of Mix Particle Size

The comparison of model performance for different types of mix particle size (i.e., seven, six, five, four, and three types of particle size) manifests that the mix particle size model was best

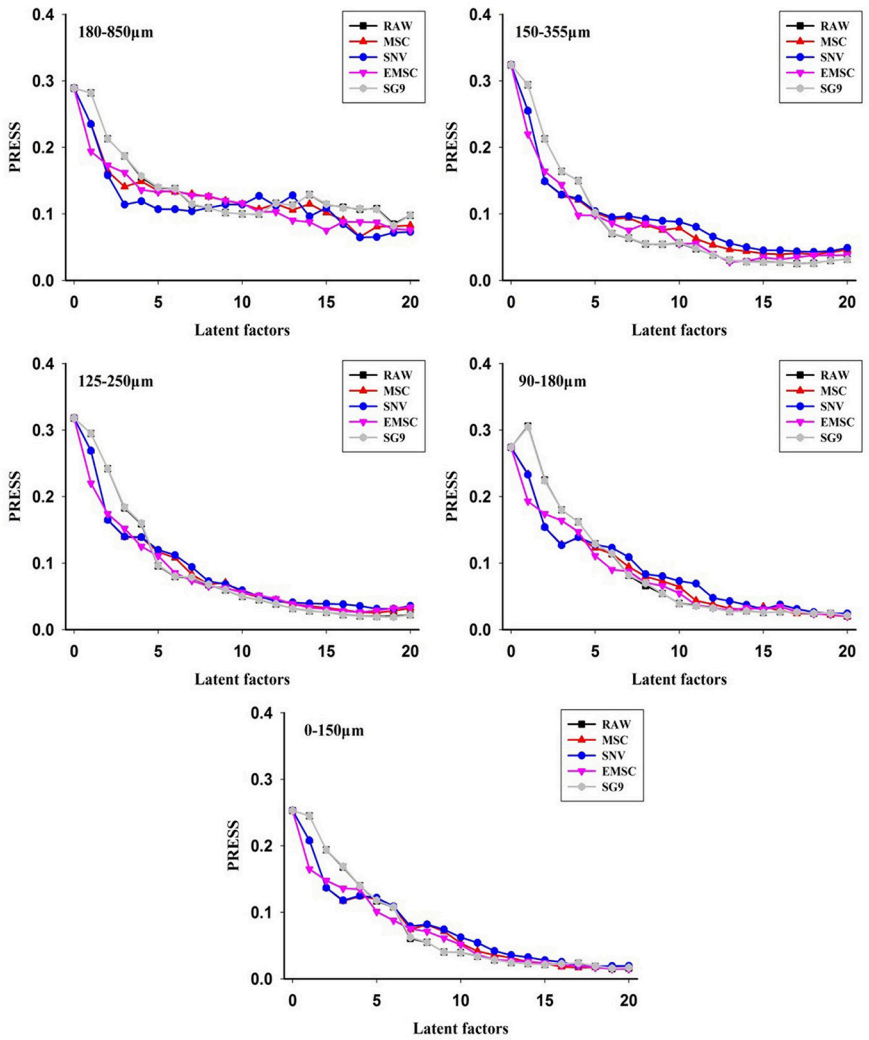



**FIGURE 6 |** The PRESS values of different preprocessing methods for mix particle size model.

constructed for 3-type mix particle size (**Table 3**). Preprocessing methods were also various, such as MSC, SNV, EMSC and SG9. It can be seen from **Figure 6**, the optimum preprocessing method for the mixed particle size of 180–850 μm, 150–355 μm, 125–250 μm, 90–180 μm, and 0–150 μm was SG9, untreated original spectra, EMSC, EMSC, and MSC, respectively, as this model has the lowest PRESS value.

The best prediction from the mix particle size model was for 90–180 μm with RPD value >3 (**Table 3**). The RPD values of other mix particle size models were also about 2, meaning that the model performance of the mix particle size models was similar. This result further revealed that particle size was vital to quantitative model performance of diffuse reflectance spectra using NIR sensor. In order to make the relationship clearer, a detailed comparison of the model of the single particle size and mixed particle size was summarized.

## Comparison of the Model Performance for Single Particle Size and Mix Particle Size

It can be concluded from the comparison between the single particle size and mixed particle size models that the RPD value of the former was better than the latter. Although the prediction results were good in the prediction performance in a certain particle size range by using a single particle size model, the prediction results of single particle size model were not stable. Most of the applications of NIR diffuse reflectance spectra were for a relatively broad range of particle sizes. As a result, a mix particle size calibration model was used for prediction in subsequent studies.

Moreover, the mix particle size correction model was also used to predict the validation set for each particle size for examining which particle size samples could be more accurately predicted as well as achieving the guideline for subsequent sample preparation. The model for particle size of 90–180 μm was selected to predict the particle size of 150–180 μm, 125–150 μm, and 90–125 μm and the best preprocessing method is MSC (**Table 4**) and RPD values of the three prediction models are 3.81, 5.78, and 2.81 (**Table 5**).

On the other hand, the RPD values of the models of single particle size were 3.40, 4.78, and 2.52. Compared with the single particle size model, the RPD value of the mix particle size model was better illustrating that the prediction of the mix particle size correction model was more accurate (**Table 5**). The relation map between the reference and validation sets was shown in **Figure 7**. The correlation between reference and prediction values was good, which further demonstrated that the mix particle size model was better than the single particle size model. Why particle size was of great importance to the quantitative model of the NIR diffuse reflectance? It was performed by the Kubelka-Munk theory, which is a critical theory in the NIR diffuse reflectance.

## Discovery of the Linear Region of NIR Diffuse Reflectance Spectra Using the Kubelka-Munk Theory

In practice, NIR diffuse reflectance is usually used for solid particle determination and its quantitative evidence is based on the Kubelka-Munk theory (**Figure 8**).

It can be learnt from the equation that the absorbance had relationship with the *k/s* value. A linear relationship was discovered between *k/s* value and *A* within a certain range.

**TABLE 4 |** The prediction model for the single particle size by using the mix particle size model.

| Mix particle size(μm) | Preprocessing | Validation (150–180) | | | Validation (125–150) | | | Validation (90–125) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | RMSEP | $R^2$ | RPD | RMSEP | $R^2$ | RPD | RMSEP | $R^2$ | RPD |
| 90–180 | RAW# | 0.2484 | 0.7945 | 2.33 | 0.1172 | 0.9369 | 4.20 | 0.1626 | 0.7839 | 2.27 |
| | MSC | 0.2109 | 0.8519 | 2.74 | 0.1723 | 0.8634 | 2.85 | 0.1670 | 0.7721 | 2.21 |
| | SNV | 0.2301 | 0.8237 | 2.51 | 0.1602 | 0.8820 | 3.07 | 0.1968 | 0.6832 | 1.87 |
| | EMSC | 0.1499 | 0.9243 | 3.81 | 0.0850 | 0.9668 | 5.78 | 0.1328 | 0.8817 | 2.81* |
| | SG9 | 0.2482 | 0.7949 | 2.33 | 0.1185 | 0.9354 | 4.15 | 0.1624 | 0.7844 | 2.27 |

*#The original spectra without any pretreatment.*

**The best prediction model for the single particle size by using the mix particle size model.*

**TABLE 5 |** Predicted results of different samples of single Scrophulariaceae Radix particle size model and calibration particle size model.

| Particle size (μm) | Single calibration model | | | Mix calibration model | | |
|---|---|---|---|---|---|---|
| | $R^2_{pre}$ | RMSEP | RPD | $R^2_{pre}$ | RMSEP | RPD |
| 150-180 | 0.9041 | 0.1696 | 3.40 | 0.9243 | 0.1499 | 3.81 |
| 125-150 | 0.9513 | 0.1029 | 4.78 | 0.9668 | 0.0850 | 5.78* |
| 90-125 | 0.8257 | 0.1460 | 2.52 | 0.8817 | 0.1328 | 2.81 |

**The best predicted results.*

As illustrated in **Figure 9**, the value for *k*/s was >4 obviously indicating that a linear region existed. This results also explained and guided the modeling performance of NIR diffuse reflectance. It was found that such a linear region provides a reference for the linear modeling of diffuse reflectance spectra. It is important to note that the linear region is beneficial for establishing a NIR diffuse reflectance model. According to our data, when the scatter coefficient *s* does not change, the absorption coefficient *k* is proportional to the sample concentration. In this study, the quantitative models for single particle size and mix particle size were both constructed to minimize the limitation that the particle size of samples was only available in a certain range. The model of single particle size was better than the mix particle size owing to a small change in the scattering coefficient *s*.

## CONCLUSIONS

Particle size is of great importance to the quantitative model of the NIR diffuse reflectance. In this study, the single particle size and mix particle size models of *Radix Scrophulariae* were constructed using PLS methods. For the single particle size model, it was obvious that the best prediction model was for

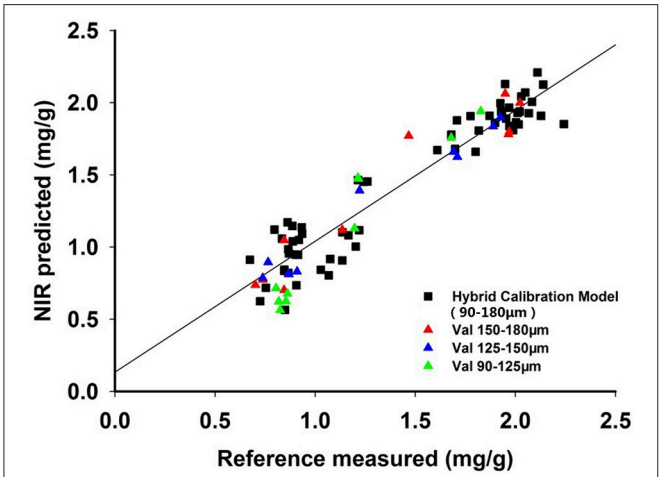



**FIGURE 7 |** The relation map of the calibration particle size models.

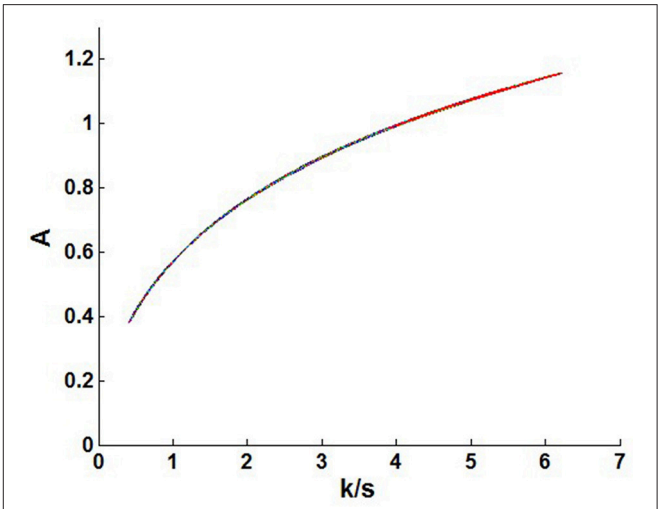



**FIGURE 9 |** The relationship between the absorbance (A) and k/s value.

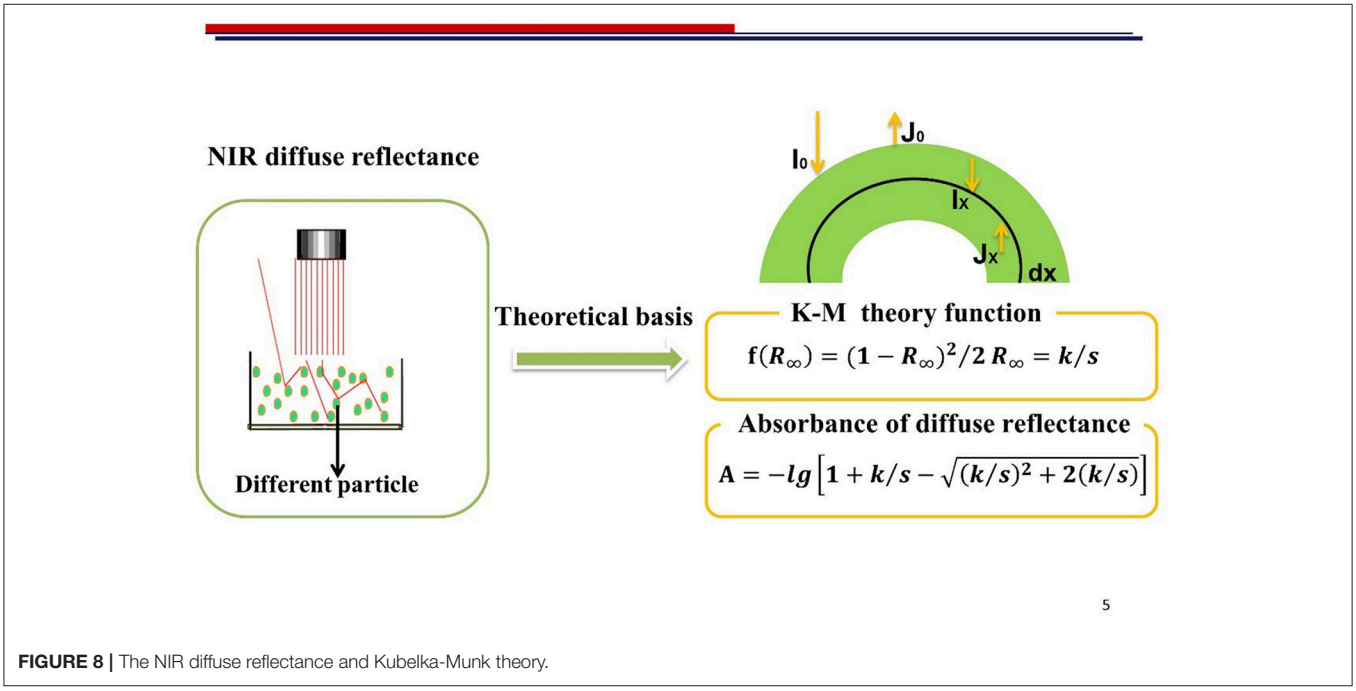



**FIGURE 8 |** The NIR diffuse reflectance and Kubelka-Munk theory.

the particle size distribution of 125–150 μm. This particle size distribution illustrated that small particle size was beneficial to construct the quantitative model of harpagoside in *Radix Scrophulariae*.

For the mix particle size model, a better prediction was obtained for the particle size distribution of 90–180 μm indicating that the mix particle size model could explain more variation in the sample, and the accuracy and robustness of the mix particle size model would be improved. Meanwhile, the quantitative evidence of NIR diffuse reflectance of different particle sizes was based on the Kubelka-Munk theory. A linear relationship was discovered between $k/s$ value and $A$ within a certain range. Data showed that a narrow range of the scatter coefficients $s$ resulted in a better model. Besides, the value for $k/s$ was >4 clearly indicating that a linear region exited. This linear region helped explain and guide the modeling performance of NIR diffuse reflectance data. Finding such a linear region provided a methodological reference for the linear modeling of NIR diffuse reflectance spectra. Thus, further accurate assessment should be obtained in advance for a precise linear model.

Our study also showed that the quantitative analysis of CHM samples was more accurate when the scattering coefficient $s$ remains unchanged or differs insignificantly at theoretical level.

## AUTHOR CONTRIBUTIONS

ZW and YQ: conceived the research; XP: performed the experiment; SD: wrote the manuscript; CD, LM, and XH: analyzed the data. All the authors prepared the manuscript and discussed the results.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81773914), Beijing Nova Program of China (xx2016050), and Science Fund for Distinguished Young Scholars in BUCM (2015-JYB-XYQ-003). The authors thank the Key Laboratory of TCM Information Engineering of State Administration of Traditional Chinese Medicine, (Beijing, China) for the assistance in data processing, Modernization of Traditional Chinese Medicine of Daozhen county of China.

## REFERENCES

Aucott, L. S., and Garthwaite, P. H. (1988). Transformations to reduce the effect of particle size in near-infrared spectra. *Analyst* 113, 1849–1854. doi: 10.1039/an9881301849

Bertelli, D., Plessi, M., Sabatini, A. G., Lollia, M., and Grillenzonib, F. (2007). Classification of Italian honeys by mid-infrared diffuse reflectance spectroscopy (DRIFTS). *Food Chem.* 101, 1565–1570. doi: 10.1016/j.foodchem.2006.04.010

Bittner, L. K., Heigl, N., Petter, C. H., Noisternig, M. F., Griesser, U. J., Bonn, G. K., et al. (2011). Near-infrared reflection spectroscopy (NIRS) as a successful tool for simultaneous identification and particle size determination of amoxicillin trihydrate. *J. Pharmaceut. Biomed.* 54, 1059–1064. doi: 10.1016/j.jpba.2010.12.019

Blanco, M., and Peguero, A. (2008). An expeditious method for determining particle size distribution by near infrared spectroscopy: comparison of PLS2 and ANN models. *Talanta* 77, 647-651. doi: 10.1016/j.talanta.2008.07.015

Borin, A., Ferrao, M. F., Mello, C., Maretto, D. A., and Poppi, R. J. (2006). Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Anal. Chim. Acta* 579, 25–32. doi: 10.1016/j.aca.2006.07.008

Bull, C. R. (1991). Compensation for particle size effects in near infrared reflectance. *Analyst* 116, 781– 786. doi: 10.1039/an9911600781

Chen, Q., Zhao, J., Liu, M., Cai, J., and Liu, J. (2008). Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms. *J. Pharm. Biomed. Anal.* 46, 568-573. doi: 10.1016/j.jpba.2007.10.031

Chen, Y., Xie, M. Y., Yan, Y., Zhu, S. B., Nie, S. P., Li, C., et al. (2008). Discrimination of *Ganoderma lucidum* according to geographical origin with near infrared diffuse reflectance spectroscopy and pattern recognition techniques. *Anal. Chim. Acta* 618, 121-130. doi: 10.1016/j.aca.2008.04.055

De Beer, T., Burggraeve, A., Fonteyne, M., Saerens, L., Remon, J. P., and Vervaet, C. (2011). Near infrared and Raman spectroscopy for the in-process monitoring ofpharmaceutical production processes. *Int. J. Pharm.* 417, 32–47. doi: 10.1016/j.ijpharm.2010.12.012

Franke, P., Gill, I., Luscombe, C. N., Rudd, D. R., Waterhouse, J., and Jayasooriya, U. A. (1998). Near-infrared mass median particle size determination of lactose monohydrate. Evaluating several chemometric approaches. *Analyst* 123, 2043–2046. doi: 10.1039/a802532k

Jin, J. W., Chen, Z. P., Li, L. M., Steponavicius, R., Thennadil, S. N., Yang, J., et al. (2012). Quantitative spectroscopic analysis of heterogeneous mixtures: the correction of multiplicative effects caused by variations in physical properties of samples. *Anal. Chem.* 84, 320-326. doi: 10.1021/ac202598f

Jing, J., Chan, C., Xu, L., Jin, D., Cao, X., Mok, D. K. W., et al. (2011). Development of an in-line HPLC fingerprint ion-trap mass spectrometric method for identification and quality control of *Radix Scrophulariae*. *J. Pharmaceut. Biomed.* 56, 830–835. doi: 10.1016/j.jpba.2011.07.032

Miyazawa, M., and Okuno, Y. (2003). Volatile components from the roots of *Scrophularia ningpoensis* Hemsl. *Flavour Frag. J.* 18, 398–400. doi: 10.1002/ffj.1232

Norris, K. H., and Williams, P. C. (1984). Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat: I. Influence of particle size. *Cereal Chem.* 61, 158–165.

Otsuka, M. (2004). Comparative particle size determination of phenacetin bulk powder by using Kubelka–Munk theory and principal component regression analysis based on near-infrared spectroscopy. *Powder Technol.* 141, 244–250. doi: 10.1016/j.powtec.2004.01.025

Pan, X. L., Li, F. Y., Wu, Z. S., Zhang, Q., Lin, Z. Z., Shi, X. Y., et al. (2015). Near infrared spectroscopy model development and variable importance in projection assignment of particle size and lobetyolin content of *Codonopsis radix*. *J. Near Infrared Spec.* 23, 327–335. doi: 10.1255/jnirs.1175

Paskatan, M. C., Steele, J. L., Spillman, C. K., and Haque, E. (2001). Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *J. Near Infrared Spectrosc.* 9, 153–164. doi: 10.1255/jnirs.303

Chinese Pharmacopoeia Commission (2010). *Chinese Pharmacopoeia Vol. 1*. Beijing: China Medical Science Press, 108.

Prieto, N., Ross, D. W., Navajas, E. A., Nute, G. R., Richardson, R. I., Hyslop, J. J., et al. (2009). On-line application of visible and near infrared reflectance spectroscopy to predict chemical-physical and sensory characteristics of beef quality. *Meat Sci.* 83, 96-103. doi: 10.1016/j.meatsci.2009.04.005

Roggo, Y., and Ulmschneider, M. (2008). "Chapter 4.3: Chemical imaging and chemometrics: useful tools for process analytical technology," in *Pharmaceutical Manufacturing Handbook: Regulations and Quality*, ed S. C. Gad (New York, NY: John Wiley & Sons), 411–431.

Sarraguça, M. C., Cruz, A. V., Amaral, H. R., Costa, P. C., and Lopes, J. A. (2011). Comparison of different chemometric and analytical methods for the prediction of particle size distribution in pharmaceutical

powders. *Anal. Bioanal. Chem*. 399, 2137-2147. doi: 10.1007/s00216-010-4230-6

The State Pharmacopoeia Commission of People's Republic of China (2015). *Pharmacopoeia of the People's Republic of China Press*. China Medical Science and Technology Press. Beijing.

Wu, Z. S., Du, M., Shi, X. Y., Xu, B., and Qiao, Y. (2015). Robust PLS prediction model for Saikosaponin A in Bupleurum chinense DC. coupled with granularity-hybrid calibration set. *J. Anal. Methods Chem*. 2015:583841. doi: 10.1155/2015/583841

Wu, Z., Ma, Q., Lin, Z., Peng, Y., Ai, L., Shi, X., et al. (2013). A novel model selection strategy using total error concept. *Talanta* 107, 248–254. doi: 10.1016/j.talanta.2012.12.057

Zornoza, R., Guerrero, C., Mataix-Solera, J., Scow, K. M., Arcenegui, V., and Mataix-Beneyto, J. (2008). Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils. *Soil Biol. Biochem*. 40, 1923-1930. doi: 10.1016/j.soilbio.2008.04.003

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 22 May 2018
doi: 10.3389/fchem.2018.00171

# Nearest Correlation-Based Input Variable Weighting for Soft-Sensor Design

*Koichi Fujiwara* and Manabu Kano*

*Department of Systems Science, Kyoto University, Kyoto, Japan*

In recent years, soft-sensors have been widely used for estimating product quality or other important variables when online analyzers are not available. In order to construct a highly accurate soft-sensor, appropriate data preprocessing is required. In particular, the selection of input variables or input features is one of the most important techniques for improving estimation performance. Fujiwara et al. proposed a variable selection method, in which variables are clustered into variable groups based on the correlation between variables by nearest correlation spectral clustering (NCSC), and each variable group is examined as to whether or not it should be used as input variables. This method is called NCSC-based variable selection (NCSC-VS). However, these NCSC-based methods have a lot of parameters to be tuned, and their joint optimization is burdensome. The present work proposes an effective input variable weighting method to be used instead of variable selection to conserve labor required for parameter tuning. The proposed method, referred to herein as NC-based variable weighting (NCVW), searches input variables that have the correlation with the output variable by using the NC method and calculates the correlation similarity between the input variables and output variable. The input variables are weighted based on the calculated correlation similarities, and the weighted input variables are used for model construction. There is only one parameter in the proposed NCVW since the NC method has one tuning parameter. Thus, it is easy for NCVW to develop a soft-sensor. The usefulness of the proposed NCVW is demonstrated through an application to calibration model design in a pharmaceutical process.

**Keywords: soft-sensor, calibration model, variable weighting, partial least squares, near infrared spectroscopy**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Daniel Cozzolino,*
*Central Queensland University, Australia*
*Larisa Lvova,*
*Università degli Studi di Roma Tor Vergata, Italy*

****Correspondence:***
*Koichi Fujiwara*
*fujiwara.koichi@i.kyoto-u.ac.jp*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *28 February 2018*
***Accepted:*** *30 April 2018*
***Published:*** *22 May 2018*

***Citation:***
*Fujiwara K and Kano M (2018) Nearest Correlation-Based Input Variable Weighting for Soft-Sensor Design. Front. Chem. 6:171. doi: 10.3389/fchem.2018.00171*

## 1. INTRODUCTION

It is important in terms of process safety and quality control to estimate product quality or other process variables, particularly when online analyzers are not available. Soft-sensors are mathematical models for estimating variables that are difficult to measure by hard sensors in real-time from other variables that are easy to measure. They have been used in various industries, for example, measurement of product composition at distillation columns in chemical processes, silicon wafer surface flatness in semiconductor processes, and active ingredient content of drugs in pharmaceutical processes. There are three methodologies for constructing soft-sensors: (i) first-principal modeling based on physicochemical knowledge of processes, (ii) statistical modeling based on process data, and (iii) a combination of the two. These methodologies also are called white-box, black-box, and gray-box modeling, respectively (Ahmad et al., 2014). In particular,

statistical modeling has attracted wide attention due to recent advances in machine learning. Although we can utilize various machine learning techniques for soft-sensor development, partial least squares (PLS) is still widely used in chemometrics as well as soft-sensor design. This is because it is possible to construct an accurate linear regression model even when the multicollinearity problem occurs (Wold et al., 2001; Kano and Ogawa, 2010; Kano and Fujiwara, 2013).

One of the major issues in developing a precise soft-sensor is input variable selection. Although soft-sensors are well-fitted to modeling data when numerous variables are used as the input, their performance may deteriorate when unimportant variables are used for estimation. In particular, input variable selection is a key when a calibration model is constructed from Near-infrared spectroscopy (NIRS) which is a powerful online measurement technology due to its short measuring time and non-invasiveness (Roggo et al., 2007; Miyano et al., 2014). The number of measured wavelengths of an NIR spectrum is usually more than 100.

If all of the possible variable combinations are tested, the computational load increases exponentially as the candidate variables increase. Appropriate variables must be selected in a systematic manner, which is referred to as input variable selection in soft-sensors, and feature selection in machine learning. A technique for input variable selection should be developed for improving the efficiency of soft-sensor design (Andersen and Bro, 2010; Mehmood et al., 2012).

In linear regression, stepwise and least absolute shrinkage and selection operator (Lasso) are widely used as input variable selection methods (Hocking, 1976; Tibshirani, 1996). In addition, PLS-Beta and variable influence on projection (VIP) are available for selecting input variables of PLS (Kubinyi, 1993).

Methods of selecting variables on the basis of correlation have been proposed because the correlation between variables should be considered when building a good regression model (Fujiwara et al., 2009). In correlation-based variable selection methods, variable groups are constructed according to the correlation, some of which are selected as the input variables. Nearest correlation spectral clustering (NCSC) (Fujiwara et al., 2010, 2011) is used for variable grouping. In NCSC-based variable selection (NCSC-VS), variable groups are constructed by NCSC, and it is examined whether or not they should be used as the input variables according to their contribution to the estimates (Fujiwara et al., 2012b). In addition, NCSC-based group Lasso (NCSC-GL) uses group Lasso (Yuan and Lin, 2006; Bach, 2008) for variable group selection after NCSC (Fujiwara and Kano, 2015). Although both NCSC-VS and NCSC-GL can build highly-accurate soft-sensors, tuning their parameters is complicated and time-consuming because they have multiple parameters to be tuned. Therefore, the number of their tuning parameters should be reduced for efficient variable selection.

Another approach is input variable weighting or input variable scaling, which multiplies each input variable by weights according to its importance from the viewpoint of estimation (Kim et al., 2014). The present work proposes an effective input variable weighting method to replace variable selection in order to conserve labor required for parameter tuning. The proposed method, referred to herein as NC-based variable weighting (NCVW), searches input variables that have the correlation with the output variable by using the NC method and calculates the correlation similarity between each input variable and the output variable. The input variables are weighted based on the calculated correlation similarities, and the weighted input variables are used for modeling. Since there is only one parameter in the proposed NCVW, an efficient soft-sensor design is realized. In this work, the usefulness of the proposed NCVW is demonstrated through application to calibration model design for estimating active pharmaceutical ingredient (API) content.

This paper is organized as follows. Section 2 introduces conventional variable selection methods for PLS modeling, and NCVW is proposed in section 3. Section 4 reports on application results of the proposed method to pharmaceutical data. The conclusion and future work are described in section 5.

## 2. CONVENTIONAL METHODS

This section introduces PLS and conventional input variable selection methods.

### 2.1. PLS

PLS is a widely used linear regression method in chemometrics as well as soft-sensor design. Given an input data matrix $\boldsymbol{X} \in \Re^{N\times M}$ whose $n$th row is the $n$th input sample $\boldsymbol{x}_n \in \Re^M$ and an output data vector $\boldsymbol{y} \in \Re^N$ whose $n$th element is the $n$th output sample $y_n \in \Re$, $\boldsymbol{X}$ and $\boldsymbol{y}$ are mean-centered and appropriately scaled. The input $\boldsymbol{X} \in \Re^{N\times M}$ and the output $\boldsymbol{y} \in \Re^N$ are broken down as follows:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{b} + \boldsymbol{f} \tag{2}$$

where $\boldsymbol{T} \in \Re^{N\times K}$ is the latent variable matrix, whose columns are the latent variable $\boldsymbol{t}_k \in \Re^N$ $(k = 1, \cdots, K)$, $\boldsymbol{P} \in \Re^{M\times K}$ is the loading matrix of $\boldsymbol{X}$ whose columns are the loading vectors $\boldsymbol{p}_k \in \Re^M$, and $\boldsymbol{b} = [b_1, \cdots, b_K]^T$ is the regression coefficient vector of $\boldsymbol{y}$. $K$ denotes the number of adopted latent variables. $\boldsymbol{E} \in \Re^{N\times M}$ and $\boldsymbol{f} \in \Re^N$ are errors.

A PLS model can be constructed by the non-linear iterative partial least squares (NIPALS) algorithm. Let the first to $k$th latent variables be $\boldsymbol{t}_1, \cdots, \boldsymbol{t}_k$, the loading vectors be $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_k$ and the loading be $b_1, \cdots, b_k$. The $(k+1)$th residual input and output are as follows:

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \boldsymbol{t}_r\boldsymbol{p}_k^T \tag{3}$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - b_k\boldsymbol{t}_k. \tag{4}$$

$\boldsymbol{t}_k$ is a linear combination of the columns of $\boldsymbol{X}_k$, that is, $\boldsymbol{t}_k = \boldsymbol{X}_k\boldsymbol{w}_k$ where $\boldsymbol{w}_k \in \Re^M$ is the $k$th weighting vector. $\boldsymbol{w}_k$ is the eigenvector corresponding the maximum eigenvalue of the following eigenvalue problem:

$$\boldsymbol{X}_{k-1}^T\boldsymbol{y}_{k-1}^T\boldsymbol{y}_{k-1}\boldsymbol{X}_{k-1}\boldsymbol{w}_k = \lambda\boldsymbol{w}_k \tag{5}$$

where $\lambda$ is an eigenvalue. The $k$th loading vector $\boldsymbol{p}_k$ and the $k$th loading $b_k$ are $\boldsymbol{p}_k = \boldsymbol{X}_k^T\boldsymbol{t}_k/\boldsymbol{t}_k^T\boldsymbol{t}_k$ and $b_k = \boldsymbol{y}_k^T\boldsymbol{t}_k/\boldsymbol{t}_k^T\boldsymbol{b}_k$.

This procedure is repeated until the number of adopted latent variables $K$ is achieved; $K$ can be determined by cross-validation.

## 2.2. PLS-Beta

PLS-Beta translates a PLS model, Equations (1, 2), into a multiple linear regression (MLR) model and selects input variables based on the magnitude of its regression coefficients (Kubinyi, 1993). The translated model is expressed as

$$\hat{\boldsymbol{y}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}\boldsymbol{y} = \boldsymbol{X}\beta_{pls} \tag{6}$$

where $\beta_{pls} = \boldsymbol{W}(\boldsymbol{P}^T\boldsymbol{W})^{-1}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{y}$, and $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K] \in \Re^{M\times K}$. The evaluation index of PLS-Beta $\nu$ is defined as

$$\nu = \frac{||\beta_{select}||}{||\beta_{pls}||} \ (0 < \nu \leq 1) \tag{7}$$

where $\beta_{select}$ is the regression coefficient vector of the selected input variables. We select individual input variables in descending order of the magnitude of $\beta_{pls}$ until $\nu$ achieves a predefined threshold.

## 2.3. Variable Influence on Projection (VIP)

The VIP evaluates the contribution of each input variable to the output (Kubinyi, 1993). The VIP score of the $j$th input variable is

$$V_j = \sqrt{M\sum_{k=1}^{K}\left(w_{jk}^2 b_k^2(\boldsymbol{t}_k^T\boldsymbol{t}_k)/||\boldsymbol{w}_k||^2\right)\Big/\sum_{k=1}^{K} b_k^2(\boldsymbol{t}_k^T\boldsymbol{t}_k)} \tag{8}$$

where $w_{jk}$ is the $j$th element of $\boldsymbol{w}_k$. Variables satisfying $V_j > \eta$ ($> 0$) are selected.

## 2.4. Stepwise

Stepwise is an input variable selection method for the MLR model based on a statistical test which checks whether or not the true value of the regression coefficient of a newly added candidate variable is zero (Hocking, 1976).

## 2.5. Least Absolute Shrinkage and Selection Operator (Lasso)

Lasso is least squares with $L_1$ regularization so that some regression coefficients approach zero (Tibshirani, 1996). The objective function of Lasso is as follows:

$$\beta_{lasso} = \arg\min_{\beta}\left(||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda||\beta||_1\right), \ \lambda \ (> 0) \tag{9}$$

Least angle regression (LARS) solves the problem of Equation (9) efficiently (Efron et al., 2004).

# 3. NEAREST CORRELATION BASED VARIABLE WEIGHTING (NCVW)

The present work proposes a new method for weighting input variables for PLS modeling to be used instead of variable selection. Since the proposed method uses the nearest correlation (NC) method for calculating correlation-based variable weights, this section explains the NC method and variable selection methods based on the NC method before the proposed method is described.

## 3.1. NC Method

The NC method was originally developed as an unsupervised learning technique for detecting samples whose correlation is similar to the query (Fujiwara et al., 2012a). The procedure of the NC method is described in Algorithm 1.

**Algorithm 1** Nearest correlation (NC) method

1: Prepare $\boldsymbol{x}_n (n = 1, \cdots N)$ and $\boldsymbol{x}_q$.
2: Set $\gamma$.
3: **for all** $n = 1, 2, \cdots, N$ $(n \neq q)$ **do**
4: $\quad \boldsymbol{x}'_n = \boldsymbol{x}_n - \boldsymbol{x}_q$.
5: **end for**
6: **for all** $k, l$ $(k \neq l)$ **do**
7: $\quad$ Calculate $C'_{k,l}$ from $\boldsymbol{x}'_k$ and $\boldsymbol{x}'_l$.
8: $\quad$ **if** $|C'_{k,l}| \geq \gamma$ **then**
9: $\quad\quad$ Output $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$ as similar samples to $\boldsymbol{x}_q$
10: $\quad$ **end if**
11: **end for**

The concept of Algorithm 1 is explained through a simple example. In **Figure 1** (left), there are seven samples $\boldsymbol{x}_q, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_6$, of which five $\boldsymbol{x}_q$ and $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_4$ are on the same plane $P$. That is, plane $P$ expresses the hidden correlation between the five samples and $\boldsymbol{x}_5$ and $\boldsymbol{x}_6$ have a different correlation. The aim of the NC method here is to detect samples whose correlation is similar to the query $\boldsymbol{x}_q$, that is, to detect $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_4$ on $P$.

In steps 3–5, the entire space is translated so that $\boldsymbol{x}_q$ becomes the origin by subtracting $\boldsymbol{x}_q$ from all other samples $\boldsymbol{x}_n$ as shown in **Figure 1** (right). The translated plane $P$ becomes the linear subspace $V$ since it contains the origin.

Draw lines connecting each sample and the origin, and check whether another sample is on the line in steps 6–8. In this example, pairs $\boldsymbol{x}_1$-$\boldsymbol{x}_4$ and $\boldsymbol{x}_2$-$\boldsymbol{x}_3$ satisfy such a relationship, and $\boldsymbol{x}_5$ and $\boldsymbol{x}_6$, which are not on $V$, cannot make pairs. At this time, the correlation coefficients of these pairs must be 1 or $-1$. Thus, the pairs whose correlation coefficients are $\pm 1$ are thought to have a correlation similar to $\boldsymbol{x}_q$. The threshold of the correlation coefficient $\gamma$ $(0 < \gamma \leq 1)$ is used for constraint relaxation. Steps 6–8 correspond to the above procedure.

Finally, the pairs whose correlations are similar to the query $\boldsymbol{x}_q$ are output in step 9.

## 3.2. NCSC

NCSC was originally proposed for sample clustering based on correlation between variables (Fujiwara et al., 2010, 2011), in which the NC method and spectral clustering (SC) (Ding et al., 2001; Ng et al., 2002) are integrated. SC is a graph theory-based clustering method, which can partition a weighted graph, whose weights express affinities between nodes, into subgraphs by cutting some of their arcs. In NCSC, the NC method is

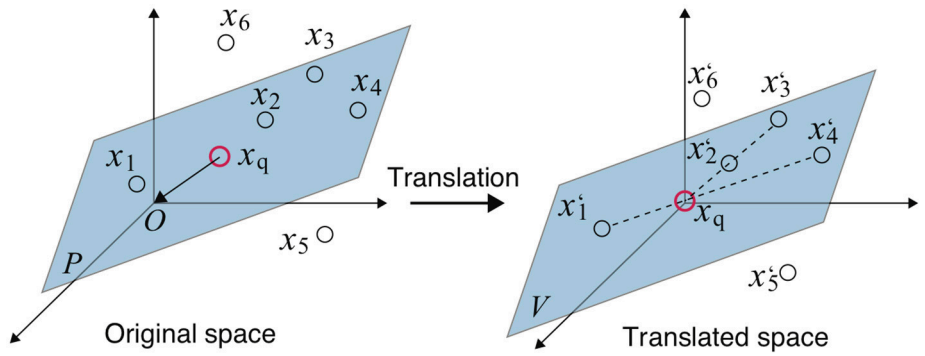



**FIGURE 1 |** Example of the NC method: A red circle is the query $\boldsymbol{x}_q$. Dashed lines denote pairs of samples, whose correlation is similar to $\boldsymbol{x}_q$ (left) (Fujiwara et al., 2010, 2012a,b).

used for building an affinity graph expressing the correlation-based similarities between samples, and SC partitions the graph constructed by the NC method.

Algorithm 2 shows an affinity matrix construction procedure in NCSC. Steps 6–13 correspond to the NC method, and the weighted graph constructed by the NC method is expressed as an affinity matrix $\boldsymbol{S}$. Although some SC algorithms have been proposed, the max-min cut (Mcut) algorithm (Ding et al., 2001) or its extended method (Ng et al., 2002) is used herein.

**Algorithm 2** Affinity matrix construction

1: Set $\gamma$ and $J$.
2: $\boldsymbol{S} \in \Re^{N\times N} \leftarrow \boldsymbol{O}_{N,N}$.
3: $L = 1$.
4: **for** $L = 1$ to $N$ **do**
5: $\quad \boldsymbol{S}_L \in \Re^{N\times N} \leftarrow \boldsymbol{O}_{N,N}$.
6: $\quad$ **for all** $n = 1, 2, \cdots, N$ $(n \neq L)$ **do**
7: $\quad\quad \boldsymbol{x}'_n = \boldsymbol{x}_n - \boldsymbol{x}_L$.
8: $\quad$ **end for**
9: $\quad$ **for all** $k, l$ $(k \neq l)$ **do**
10: $\quad\quad$ Calculate $C'_{k,l}$ from $\boldsymbol{x}'_k$ and $\boldsymbol{x}'_l$.
11: $\quad\quad$ **if** $|C'_{k,l}| \geq \gamma$ **then**
12: $\quad\quad\quad (\boldsymbol{S}_L)_{k,l} = (\boldsymbol{S}_L)_{l,k} = 1$.
13: $\quad\quad$ **end if**
14: $\quad$ **end for**
15: $\quad \boldsymbol{S} = \boldsymbol{S} + \boldsymbol{S}_L$.
16: **end for**

NCSC has two parameters: the threshold in the NC method $\gamma$ and the number of clusters partitioned by SC, $J$. Previous studies have suggested the default value of $\gamma$ to be 0.99 (Fujiwara et al., 2010, 2011), and that $J$ needs to be determined by trial and error.

## 3.3. NCSC-VS and NCSC-GL

NCSC has been utilized for variable selection in soft-sensor design. In these methods, multiple variable groups are constructed by NCSC, of which some are selected as the input variables of a soft-sensor. NCSC classifies variables into $J$ variable groups $\boldsymbol{v}_j = \{x_m \mid m \subset \mathcal{V}_j\}$ $(j = 1, \cdots, J)$, where $\mathcal{V}_j$ is the subset of variable indexes and $\mathcal{V} = \cup\mathcal{V}_j$. An affinity matrix is derived from the transposed input variable matrix $\boldsymbol{X}^T$ by the NC method for variable grouping.

NCSC-VS evaluates each variable group as to whether or not its members should be used as input variables from the viewpoint of contribution to the output (Fujiwara et al., 2012b). The $j$th PLS model with the number of latent variables $P$, $f_j^P$, is built from the $j$th variable group matrix $\boldsymbol{X}_j$, and its contribution is evaluated by

$$C_j^P = 1 - \frac{||\hat{\boldsymbol{y}}_j^P||^2}{||\boldsymbol{y}||^2} \tag{10}$$

where $\hat{\boldsymbol{y}}_j^P$ is the estimate of $f_j^P$. We select $D$ $(\leq J)$ variable groups in descending order of $C_j^P$ and construct the final PLS model from the selected input variables.

NCSC-GL selects variable groups by using group Lasso instead of contribution evaluation in NCSC-VS. Group Lasso is an extension of Lasso for selecting some input variable groups from predefined multiple variable groups (Yuan and Lin, 2006; Bach, 2008).

Suppose that $M$ variables are divided into $J$ groups; and $\boldsymbol{X}_j$ and $\beta_j$ denote the input data matrix and the regression coefficient vector corresponding to the $j$th group, respectively. The number of variables in the $j$th group is $M_j$, that is, $M = \sum_{j=1}^{J} M_j$. The regression coefficients of group Lasso is derived as:

$$\beta_{glasso} = \arg\min_{\beta} \left( ||\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{X}_j\beta_j||_2^2 + \lambda \sum_{j=1}^{J} \sqrt{M_j}||\beta_j||_2 \right) \tag{11}$$

where $\beta = [\beta_1^T, \cdots, \beta_J^T]^T$, and $\lambda$ is a parameter. Variable groups must be constructed in advance in group Lasso. Thus, NCSC-GL uses variable groups formed by NCSC as the input of group Lasso.

NCSC-VS has four tuning parameters: $\gamma$ in the NC method, the number of variable groups partitioned by SC, $J$, latent variables in the PLS models for variable group evaluation, $P$, and selected variable groups, $D$. On the other hand, there are three tuning parameters in NCSC-GL: $\gamma$ in the NC method, the number of variable groups $J$ formed by SC and $\lambda$ in group Lasso. These three or four parameters need to be tuned for appropriate

input variable selection. However, their joint optimization is burdensome and time-consuming. For more efficient soft-sensor design, the number of tuning parameters should be reduced.

## 3.4. NCVW

A new input variable weighting method, referred to as NC-based variable weighting (NCVW), is proposed to be used instead of variable selection for conserving labor required for parameter tuning. The proposed method applies the NC method to the input variables and output variable together for calculating similarities based on the correlation between the input variables and output variable, and uses the input variables weighted by the calculated similarities for modeling.

Let the $n$th input sample and the $n$th output sample are $\boldsymbol{x}_n \in \Re^M$ and $y_n$, where $M$ denotes the number of input variables. In NCVW, the NC method is applied to extended samples

$$\boldsymbol{x}'_n = [x_n^{[1]}, \cdots, x_n^{[M]}, y_n]^T \ (n = 1, \cdots, N) \tag{12}$$

and the affinity matrix $\boldsymbol{S}'$ is constructed. Next, the 1st to $M$th element in the $(M+1)$th column of $\boldsymbol{S}$ which corresponds to the output variable is extracted as a weighting vector $\boldsymbol{w} = [w^{[1]}, \cdots, w^{[M]}]$. Finally, a new input variable for PLS modeling is formed as

$$\boldsymbol{z}_n = \boldsymbol{w} \circ \boldsymbol{x} = [w^{[1]}x^{[1]}, \cdots, w^{[M]}x^{[M]}]^T. \tag{13}$$

where $\boldsymbol{a} \circ \boldsymbol{b}$ denotes an element-wise product between vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Algorithm 3 summarizes the procedure of the proposed NCVW.

**Algorithm 3** Nearest correlation based variable weighting (NCVW)

1: Prepare $\boldsymbol{x}_n$ and $y_n$ $(n = 1, \cdots N)$.
2: $\boldsymbol{x}_n \longleftarrow [x_n^{[1]}, \cdots, x_n^{[M]}, y_n]^T \ (n = 1, \cdots, N)$
3: Get $\boldsymbol{S} \in \Re^{(M+1)\times(M+1)}$ by applying Algorithm 2 to $\boldsymbol{x}_n$.
4: Extract the 1st to $M$th element in the $M+1$th column of $\boldsymbol{S}$ as $\boldsymbol{w} = [w^{[1]}, \cdots, w^{[M]}]$.
5: $\boldsymbol{z}_n = \boldsymbol{w} \circ \boldsymbol{x} = [w^{[1]}x^{[1]}, \cdots, w^{[M]}x^{[M]}]^T \ (n = 1, \cdots N)$.
6: Construct a model from $\boldsymbol{z}_n$ by PLS.

In soft-sensor design, the correlation among multiple input variables needs to be considered as well as the correlation between an individual input variable and the output variable. Thus, the proposed NCVW does not evaluate the correlation between each input variable and the output variable, but the correlation of multiple input variables together, which may contribute to an improvement in the estimation performance of a soft-sensor. In addition, the proposed NCVW has only one parameter, which is the threshold of the NC method $\gamma$. This leads to a huge efficiency improvement of soft sensor development.

# 4. CASE STUDY

This case study evaluates the performance of the proposed NCVW through application to pharmaceutical data provided by Daiichi Sankyo Co., Ltd. (Kim et al., 2011).

## 4.1. Objective Data

The objective of this case study is to design a calibration model that estimates active pharmaceutical ingredient (API) content in a target drug. NIR spectra (2203 points in 800−2500 nm) and the API content were measured from the granules of the drug through experiments. Since the number of wavelengths in NIR spectra was large, appropriate input wavelengths of NIR spectra had to be selected for constructing a precise calibration model. The modeling data and validation data consisted of 576 and 20 samples, respectively.

## 4.2. Model Construction

Before modeling, a first-order differential Savitzky-Golay smoothing filter (Savitzky and Golay, 1964) was applied to the spectra. As a benchmark, a PLS model using all the wavelengths as the input was constructed, which was called PLS-All. The number of its adopted latent variables was determined by cross-validation. Input wavelengths were selected using PLS-Beta, VIP, stepwise, Lasso, NCSC-VS, and NCSC-GL. Parameters used in each method were selected by trial and error, which are shown in **Table 1**. We calculated the root-mean-square error (RMSE) for the modeling data in each parameter and determined the optimal wavelengths based on the calculated RMSE.

We designed PLS models with the wavelengths selected by each method in which cross-validation was used for determining the appropriate number of latent variables. Although Lasso derives regression coefficients, the PLS model was built from the wavelengths whose regression coefficient was not zero. This is for the reason that the number of retained wavelengths was still large and dimension reduction by PLS may have been needed. On the other hand, in the proposed NCVW, we calculated variable weights and constructed the PLS model from the weighted wavelengths. Finally, the API content was estimated by these constructed PLS models.

These procedures were repeated 100 times for calculating average CPU time per one modeling of each method. The computer configuration was as follows: OS: Windows10 (64bit),

**TABLE 1** | Tested parameters.

| | **Parameters** |
|---|---|
| PLS-All | – |
| PLS-Beta | $\nu = \{0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$ |
| VIP | $\eta = \{0.6, 0.7, 0.8, 0.9, 1.0, 1.1\}$ |
| Lasso | $\lambda = \{0.1, 0.2, 0.4, 0.5, 0.8, 1.0\}$ |
| Stepwise | $\bar{p} = \{0.005, 0.05, 0.08, 0.1, 0.12, 0.15\}$ |
| NCSC-VS | $\gamma = 0.99$ |
| | $J = \{5, 6, 7, 8, 9, 10\}$ |
| | $P = \{9, 10, 11\}$ |
| | $D = \{2, 3\}$ |
| NCSC-GL | $\gamma = 0.99$ |
| | $J = \{5, 6, 7, 8, 9, 10\}$ |
| | $\lambda = \{20, 25\}$ |
| NCVW | $\gamma = 0.99$ |

CPU: Intel Core i7-8700 (3.2 GHz×6), RAM: 64G bytes, and MATLAB 2018a.

**Table 2** summarizes the results of the case study. #Wavelength and #LV mean the numbers of selected wavelengths and adopted latent variables determined by cross-validation, $R^2$ is the determination coefficient, "CPU time" is the average CPU times [s], and "Parameters" denotes the optimal parameters in each method. In addition, **Figure 2** shows the detailed estimation results.

**TABLE 2 |** API content estimation results.

| | #WL | #LV | Parameters | RMSE | $R^2$ | CPU time [s] |
|---|---|---|---|---|---|---|
| PLS-All | 2203 | 37 | – | 1.28 | 0.83 | – |
| PLS-Beta | 928 | 36 | $\nu = 0.75$ | 1.06 | 0.81 | 1.52 |
| VIP | 1133 | 19 | $\eta = 0.8$ | 1.01 | 0.83 | 0.36 |
| Lasso | 1138 | 39 | $\lambda = 0.2$ | 0.98 | 0.87 | 0.17 |
| stepwise | 561 | 24 | $\tilde{p} = 0.15$ | 1.42 | 0.72 | 1.64 |
| NCSC-VS | 843 | 25 | $\gamma = 0.99, J = 6, P = 10, D = 2$ | 0.77 | 0.92 | 202.39 |
| NCSC-GL | 1059 | 18 | $\gamma = 0.99, J = 8, \lambda = 25$ | 0.71 | 0.93 | 204.04 |
| NCVW | 2203 | 15 | $\gamma = 0.99$ | 0.74 | 0.92 | 202.27 |

While PLS-Beta, VIP, and Lasso improved the estimation performance compared to PLS-All, only stepwise was worse than PLS-All. Both NCSC-VS and NCSC-GL achieved higher performance than methods above; and, in particular, NCSC-GL had the best performance. The proposed NCVW achieved almost the same performance as NCSC-VS and NVSC-GL, even though NCVW has only one tuning parameter. RMSE of NCVW was improved by about 42% in comparison with PLS-All.

It is concluded that the proposed NCVW is a tuning-free soft-sensor design technique and that its performance is comparable to the NCSC-based methods.

## 4.3. Discussion

According to **Table 2**, the CPU time of NCSC-VS, NCSC-GL, and the proposed NCVW were much longer than those of other methods. NCSC occupied more than 99% of their CPU time since it uses iteration for similarity calculation, which means NCVW does not improve the computational load. In addition, the estimation performance of NCVW was not improved in

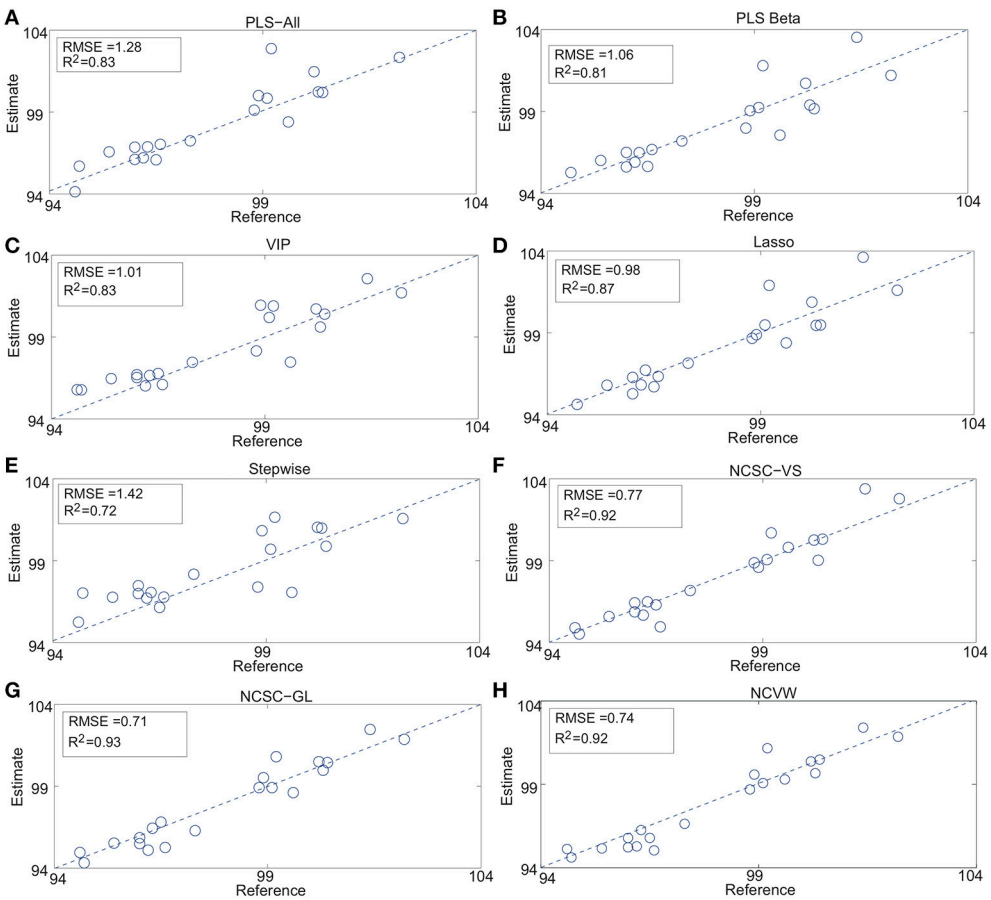



**FIGURE 2 |** API content estimation results: **(A)** PLS-All, **(B)** PLS Beta, **(C)** VIP, **(D)** Lasso, **(E)** Stepwise, **(F)** NCSC-VS, **(G)** NCSC-GL and, **(H)** NCVW (Fujiwara et al., 2010, 2012a,b).

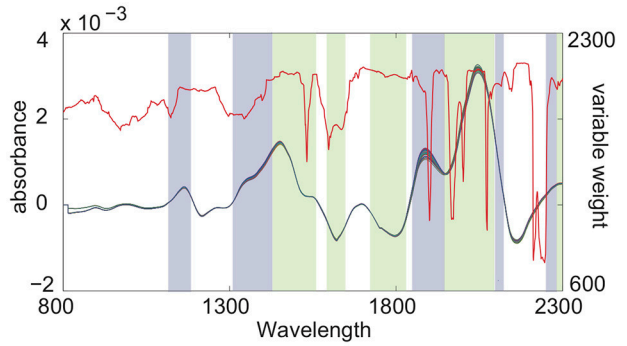



**FIGURE 3 |** Wavelength group selection by NCSC-VS and variable weights by NCVW.

comparison with NCSC-GL; however, construction of the actual soft-sensor therewith is much easier than NCSC-VS and NCSC-GL. The latter methods respectively have four and three tuning parameters. In this case study, 36 calculations in NCSC-VS and 12 calculations in NCSC-GL were repeated for searching the best parameter combination according to **Table 1**. It becomes difficult to find the optimal parameter combination when the number of tuning parameters increases. On the other hand, NCVW has just one parameter–the threshold of the NC method $\gamma$ and its recommended value has been proposed to be $\gamma = 0.99$ (Fujiwara et al., 2010, 2011). In fact, the total computation times of NCSC-VS, NCSC-GL, and the proposed NCVW were about 121, 42, and 3 min, respectively, for parameter tuning in this case study. Thus, the proposed NCVW makes the soft-sensor design much more efficient than NCSC-VS and NCSC-GL.

Variable weighting based on another type of the weight, the correlation coefficient between each input variable and the output variable, was evaluated. This method is called correlation coefficient-based variable weighting (CCVW). The $m$th variable weight of CCVW is defined as follows:

$$c^{[m]} = \frac{\boldsymbol{y}^T \boldsymbol{x}^{[m]}}{||\boldsymbol{y}||||\boldsymbol{x}^{[m]}||} \tag{14}$$

where $\boldsymbol{x}^{[m]} \in \Re^N$ denotes the $m$th column in the input data matrix $\boldsymbol{X} \in \Re^{N \times M}$ and $\boldsymbol{y} \in \Re^N$ is the output data vector. A PLS model was constructed from the input variables weighted by $c^{[m]}$. RMSE and $R^2$ of NCVW were 1.34 and 0.84, respectively. This showed the effectiveness of the variable weight by NCVW which consider the correlation of multiple input variables and the output variable together.

**Figure 3** shows the results of wavelength selection of NCSC-VS and the variable weights calculated by the proposed NCVW. The colored bands express the selected wavelengths, and the colors denote groups by NCSC-VS. The red line is the weights of NCVW. The wavelength groups selected by NCSC-VS contained almost only specific peaks. On the other hand, in NCVW, the weights of almost all wavelength regions that contain peaks, were large while some peaks had small weights. This is consistent with the physicochemical knowledge that information about compounds is contained in specific peaks. Some peaks might have important information about the API content, and other peaks might not contribute to API content estimation. Therefore, the weights by NCVW suggest that unnecessary peaks for API content estimation exist in NIR spectra. This indicates that NCVW can create meaningful weights for soft-sensor design.

## 5. CONCLUSION

In the present work, an input variable weighting method was proposed for efficient and highly-accurate soft-sensor design. The proposed NCVW derives the variable weights on the basis of the correlation between the input variables and output variable by utilizing the NC method and builds a PLS model from the weighted input variables. Since NCVW has just one tuning parameter, its soft-sensor design is efficient. The performance of NCVW was evaluated through the case study of calibration model development of the pharmaceutical process. The result showed that the estimation performance of NCVW was comparable to that of NCSC-VS and NCSC-GL, while the labor required for parameter tuning was greatly conserved. Although the objective data used in the case study was NIR spectra data, the application area of the proposed method is not limited to a specific type of data. The proposed NCVW is applicable to general soft-sensor design when the number of input variables is large. Therefore, NCVW will contribute to realizing the efficient soft-sensor design.

## AUTHOR CONTRIBUTIONS

KF developed the proposed method, analyzed the data, and wrote the initial draft of the manuscript. MK contributed to data collection and analysis and assisted in the preparation of the manuscript. Both authors approved the final version of the manuscript, and agree to be accountable for all aspects of the work.

## FUNDING

This work was partially supported by the JFE 21st Century Foundation.

## ACKNOWLEDGMENTS

The authors thank Daiichi-Sankyo Co., Ltd. for providing real operation data used in case studies.

## REFERENCES


Ahmad, I., Kano, M., Hasebe, S., Kitada, H., and Murata N. (2014). Gray-box modeling for prediction and control of molten steel temperature in tundish. *J. Process Control* 24, 375–382. doi: 10.1016/j.jprocont.2014.01.018

Andersen, C. M., and Bro, R. (2010). Variable selection in regression – a tutorial. *J. Chemometrics* 24, 728–737. doi: 10.1002/cem.1360

Bach, F. (2008). Consistency of group lasso and multiple kernel learning. *J. Mach. Learn. Res.* 9, 1179–1225.

Ding, C. H. Q., He, X., Zha, H., Gu, M., and Simon, H. D. (2001). "A min-max cut algorithm for graph partitioning and data clustering," in *IEEE International Conference on Data Mining (ICDM)* (San Jose, CA) 107–114.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* 32, 407–499. doi: 10.1214/009053604000000067

Fujiwara, K., and Kano, M. (2015). Efficient input variable selection for soft-senor design based on nearest correlation spectral clustering and group lasso. *ISA Trans.* 58, 367–379. doi: 10.1016/j.isatra.2015.04.007

Fujiwara, K., Kano, M., and Hasebe, S. (2009). Soft-sensor development using correlation-based just-in-time modeling. *AIChE J.* 55, 1754–1765. doi: 10.1002/aic.11791

Fujiwara, K., Kano, M., and Hasebe, S. (2010). Development of correlation-based clustering method and its application to software sensing. *Chemom. Intell. Lab. Syst.* 101, 130–138. doi: 10.1016/j.chemolab.2010.02.006

Fujiwara, K., Kano, M., and Hasebe, S. (2012a). Development of correlation-based pattern recognition algorithm and adaptive soft-sensor design. *Control Eng. Pract.* 20, 371–378. doi: 10.1016/j.conengprac.2010.11.013

Fujiwara, K., Kano, M., and S.Hasebe (2011). Correlation-based spectral clustering for flexible process monitoring. *J. Process Control* 21, 1348–1448. doi: 10.1016/j.jprocont.2011.06.023

Fujiwara, K., Sawada, H., and Kano, M. (2012b). Input variable selection for pls modeling using nearest correlation spectral clustering. *Chemom. Intell. Lab. Syst.* 118, 109–119. doi: 10.1016/j.chemolab.2012.08.007

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* 32, 1–49.

Kano, M., and Fujiwara, K. (2013). Virtual sensing technology in process industries: trends and challenges revealed by recent industrial applications. *J. Chem. Eng. Jpn.* 46, 1–17. doi: 10.1252/jcej.12we167

Kano, M., and Ogawa, M. (2010). The state of the art in chemical process control in japan: Good practice and questionnaire survey. *J. Process Control* 20, 969–982. doi: 10.1016/j.jprocont.2010.06.013

Kim, S., Kano, M., Nakagawa, H., and Hasebe, S. (2014). Input variable scaling for statistical modeling. *Comput. Chem. Eng.* 74, 59–65. doi: 10.1016/j.compchemeng.2014.12.016

Kim, S., Kano, M., Nakagawa, H., and Hasebe, S. (2011). Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. *Int. J. Pharm.* 421, 269–274. doi: 10.1016/j.ijpharm.2011.10.007

Kubinyi, H. (1993). *3D QSAR in Drug Design; Theory, Methods, and Applications*. Leiden; Holland: ESCOM.

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemom. Intell. Lab. Syst.* 118, 62–69. doi: 10.1016/j.chemolab.2012.07.010

Miyano, T., Kano, M., Tanabe, H., Nakagawa, H., Watanabe, T., and Minami, H. (2014). Spectral fluctuation dividing for efficient wavenumber selection: application to estimation of water and drug content in granules using near infrared spectroscopy. *Int. J. Pharm.* 475, 504–513. doi: 10.1016/j.ijpharm.2014.09.007

Ng, A. N., Jordan, M. I., and Weiss, Y. (2002). On "spectral clustering: Analysis and an algorithm," in *NIPS* (Vancouver, BC), 849–856.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., and Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.* 44, 683–700. doi: 10.1016/j.jpba.2007.03.023

Savitzky, A., and Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 627–1639.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* 58, 267–288.

Wold, S., Sjostroma, M., and Erikssonb, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* 58, 109–130. doi: 10.1016/S0169-7439(01)00155-1

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x


**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 24 May 2018
doi: 10.3389/fchem.2018.00184

# Compilation of a Near-Infrared Library for Construction of Quantitative Models of Oral Dosage Forms for Amoxicillin and Potassium Clavulanate

*Wen-bo Zou, Xiao-meng Chong, Yan Wang and Chang-qin Hu**

*Antibiotic Division, National Institutes for Food and Drug Control, Beijing, China*

The accuracy of quantitative models for near-infrared (NIR) spectroscopy is dependent upon calibration samples with concentration variations. Conventional sample-collection methods have shortcomings (especially time-consumption), which creates a "bottleneck" in the application of NIR models for Process Analytical Technology (PAT) control. We undertook a study to solve the problem of sample collection for construction of NIR quantitative models. Amoxicillin and potassium clavulanate oral dosage forms (ODFs) were used as examples. The aim of this study was to find an approach to construct NIR quantitative models rapidly using a NIR spectral library based on the idea of a universal model. The NIR spectral library of amoxicillin and potassium clavulanate ODFs was defined and comprised the spectra of 377 batches of samples produced by 26 domestic pharmaceutical companies, including tablets, dispersible tablets, chewable tablets, oral suspensions, and granules. The correlation coefficient ($r_T$) was used to indicate the similarities of the spectra. The calibration sets of samples were selected from a spectral library according to the median $r_T$ of the samples to be analyzed. The $r_T$ of the samples selected was close to the median $r_T$. The difference in $r_T$ of these samples was 1.0–1.5%. We concluded that sample selection was not a problem when constructing NIR quantitative models using a spectral library compared with conventional methods of determining universal models. Sample spectra with a suitable concentration range in NIR models were collected rapidly. In addition, the models constructed through this method were targeted readily.

**Keywords: near-infrared spectroscopy, universal model, sample selection, spectral library, quantitative analysis**

OPEN ACCESS

***Edited by:***
*Federico Marini,*
*Sapienza Università di Roma, Italy*

***Reviewed by:***
*Thiagarajan Soundappan,*
*Navajo Technical University, United States*
*Huawen Wu,*
*BaySpec, Inc., United States*

****Correspondence:***
*Chang-qin Hu*
*hucq@nifdc.org.cn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *02 November 2017*
***Accepted:*** *07 May 2018*
***Published:*** *24 May 2018*

***Citation:***
*Zou W, Chong X, Wang Y and Hu C (2018) Compilation of a Near-Infrared Library for Construction of Quantitative Models of Oral Dosage Forms for Amoxicillin and Potassium Clavulanate. Front. Chem. 6:184. doi: 10.3389/fchem.2018.00184*

## INTRODUCTION

Near infrared spectroscopy (NIRS) is a rapid, low-cost, and non-destructive technology that has been used widely in quality control and for the rapid detection of pharmaceuticals (Jamrógiewicz, 2012; Chong et al., 2016; Dong et al., 2016). It has also been used to monitor pharmaceutical manufacturing online (Möltgen et al., 2012; Sarraguça et al., 2014; Wahl et al., 2014). In 2003, US The

Food and Drug Administration (FDA) announced Pharmaceutical Current Good Manufacturing Practices (cGMPs) for the twenty-first century to obtain better knowledge of production processes. The document offers guidelines to secure pharmaceutical quality *via* process control of raw materials as well as intermediate and final products during manufacturing (Velagaleti et al., 2002; United States Food and Drug Administration, 2004). Process Analytical Technology (PAT) is the key point of process control during pharmaceutical production (United States Pharmacopeial Convention, 2015). NIRS is the most frequently used method of PAT because it is efficient, pollution-free and has no need for sample pretreatment (Hertrampf et al., 2015).

The accuracy of quantitative analysis depends on NIR models. Sample selection is challenged during the selection of NIR quantitative models. A sufficient number of samples are needed to comprise the appropriate concentration range necessary for the calibration set. However, collecting enough calibration samples with concentration variability in the PAT process is difficult.

Five methods have been proposed to collect calibration samples. The first method uses normal products and the development of samples, which are normally out of specification and can extend the concentration range (Gottfries et al., 1996; Merckle and Kovar, 1998; Corti et al., 1999). The second method uses standard additions for active pharmaceutical ingredients (APIs) or excipients to increase or decrease the sample concentration (Dreassi et al., 1996; Blanco et al., 1997, 2001). The third method uses laboratory-made samples by changing the concentration of the components in the matrix (Moffat et al., 2000; Blanco et al., 2001). The fourth method uses laboratory-made samples with production samples that comprise granules, tablet cores, and coated tablets (these are all sources of variation in the model) (Blanco et al., 1998). The fifth method uses a mixture of API and excipients in different proportions for preparation of laboratory-scale samples (Mafalda and Lopes, 2009).

These methods can broaden the range of the calibration concentration. However, the sample-preparation procedure is time-consuming. Also, the samples prepared in the laboratory are not "real" commercial products because they cannot encompass all the chemical and physical properties of commercial products (e.g., excipients, particle size, polymorphs). Besides, constructing models using underdosed and overdosed samples may carry problems in terms of the correlation between the concentrations of API and other excipients (Mafalda and Lopes, 2009). When constructing models of compound preparations, the underdosing/overdosing procedure should be done by means of a "sample concentration matrix." This involves calculation of the cross-correlation between the constituents as their individual concentrations are increased or decreased, thereby avoiding spurious correlations among constituents (Blanco and Alcala, 2006). Therefore, sample selection remains a "bottleneck" in the application of NIR models for PAT control.

We have been studying NIR universal models (Feng et al., 2010). Such a universal model could be used to rapidly analyze pharmaceuticals from different manufacturers under the same international non-proprietary name (INN). A homologous sample based on the application of universal samples has been proposed (Zou et al., 2013). A set of samples are considered "homologous" if they contain the same API, similar excipients, and similar production processes. The NIR spectra of the samples in one homologous sample set are, therefore, highly similar. Calibration sets in the universal model comprise several homologous samples. Samples can be accurately analyzed *via* universal models if they fall into homologous samples from the calibration set. Errors may occur, and the original model should be updated if the universal model analyzes a new sample that cannot be covered by the existing homologous sample sets. Universal models do not need sample preparation. All of the calibration and validation samples can be obtained in the market. The method of sample selection ensures an appropriate range of calibration concentration, which is important to develop a robust calibration.

Amoxicillin and potassium clavulanate are compound preparations of β-lactam and β-lactamase inhibitors, respectively. They are used for the treatment of bacterial infections of the respiratory and urinary tracts. The oral dosage forms (ODFs) for amoxicillin and potassium clavulanate combined in different ratios are tablets (7:1, 4:1, 2:1), dispersible tablets (14:1, 7:1, 4:1), chewable tablets (8:1, 2:1), granules (7:1, 4:1), and oral suspensions (7:1, 4:1, 2:1). Universal models of tablets of amoxicillin and potassium clavulanate are constructed to measure the content of amoxicillin, potassium clavulanate, water, and the major impurity: cycle-closed dimer (Chong et al., 2016). Some NIR methods have been proposed for determination of amoxicillin in suspensions and capsules, in which calibration samples are formulated similar to those for commercial products (Silva et al., 2012; Khan et al., 2016).

Herein, we took the concept of a universal model to build a NIR spectral library of ODFs for amoxicillin and potassium clavulanate by collecting various products with different strengths from different manufacturers. Calibration samples could be chosen from the NIR spectral library when establishing NIR universal models to determine the contents of amoxicillin, potassium clavulanate, and/or water in the PAT control. Samples were considered to be homologous if they were similar to calibration samples. The feasibility of constructing NIR models using a NIR spectral library was discussed. Thus, the problem of collecting calibration samples could be resolved by PAT control.

# MATERIALS AND METHODS

## Samples and Reagents

Three hundred and seventy seven batches of amoxicillin and potassium clavulanate ODFs produced by 26 manufacturers were collected in post-marketing surveillance in 2012 and 2014. There were 74 batches of tablets, 78 batches of dispersible tablets, 10 batches of chewable tablets, 96 batches of granules, and 120 batches of oral suspensions; 211 samples of amoxicillin capsule were from 100 batches provided by ZhuHai United Laboratories. The amoxicillin capsules included mixed intermediate granules of amoxicillin capsules as well as filled capsules and/or packaged

capsules of the same batch. A reference standard of amoxicillin trihydrate (lot number: 130409-201011; content: 85.8%) and potassium clavulanate (lot number: 130429-201307; content: 95.0%) were provided by the US National Institutes for Food and Drug Control.

Methanol was purchased from Fisher Scientific (Pittsburgh, PA, USA). Phosphoric acid was obtained from Beijing Chemical Works (Beijing, China). Sodium dihydrogen phosphate dihydrate was purchased from Sinopharm Chemical Reagents (Beijing, China).

## Reference Method

The reference contents of amoxicillin and potassium clavulanate were determined by high-performance liquid chromatography (HPLC) (Chong et al., 2016) using an Ultimate 3000 HPLC system (Dionex, Sunnyvale, CA, USA) and an ZORBAX SB-C18 column (5 μm, 150 × 4.6 mm; Agilent Technologies, Santa Clara, CA, USA). The chromatographic conditions were: column temperature, 30°C; detection wavelength, 220 nm; flow rate, 1 mL $min^{-1}$; injection volume, 20 μL; mobile phase, 5:95 (v/v) methanol/phosphate buffer (0.05 mol $L^{-1}$ sodium dihydrogen phosphate pH adjusted to 4.4 with 10% phosphoric acid).

For each tablet or granule/oral suspension of amoxicillin and potassium clavulanate, 10 tablets or 10 bags of granules/oral suspensions were pulverized in a motor, weighed accurately, dissolved in the mobile phase to get 0.5 mg $mL^{-1}$ of amoxicillin or potassium clavulanate for HPLC analysis. Two replicate runs were done for each sample to get the average reference value. The water content was determined *via* the Karl Fischer method according to the *Chinese Pharmacopoeia*.[1]

## Acquisition and Pre-processing of NIR Spectra

Acquisition of NIR spectra was done on a MATRIX-F FT-NIR spectrometer (Bruker Optics, Billerica, MA, USA) equipped with a 1.5-mm fiberoptic diffuse reflectance probe and an extended TE-cooled indium gallium arsenide (InGaAs) detector. Data were collected and processed using OPUS v6.5 software (Bruker Optics).

The fiberoptic probe was used to record diffuse reflectance spectra at 8 $cm^{-1}$ resolution in the spectral range 4,000–12,000 $cm^{-1}$. During each measurement, 32 co-added scans were undertaken. The measurement was carried out by putting the fiberoptic diffuse reflectance probe close to the sample. For each tablet, dispersible tablet, and chewable tablet of amoxicillin and potassium clavulanate, three tablets were selected randomly and measured. The weight of each tablet was 0.5–1.0 g. Three sample bags, weighing 3.0–6.0 g, were selected randomly and measured for a granule and oral suspension of amoxicillin and potassium clavulanate. For each mixed intermediate granule of an amoxicillin capsule, 5 g of powder was placed in a vial and measured in triplicate. For each filled capsule and packaged capsule of amoxicillin, 5 g of powder of the capsule was placed in a vial and measured thrice. The three original spectra were averaged by OPUS v6.5 software. The average spectra were then subjected to a Savitzky–Golay first derivative treatment with 17-point smoothing, followed by vector normalization transformation. The pre-processed spectra were used for construction and validation of the model.

## Compilation of a NIR Spectral Library

The NIR spectral library comprised the spectra of 377 batches of amoxicillin and potassium clavulanate ODFs produced by 26 manufacturers (74 batches of tablets, 78 batches of dispersible tablets, 10 batches of chewable tablets, 96 batches of granules, and 120 batches of oral suspensions). For the NIR spectra of the library, the content of amoxicillin was 4.77–57.86%, the content of potassium clavulanate was 1.03–20.17%, and the water content was 0.24–9.30%. The correlation coefficient $r_T$ between the spectra of each amoxicillin and potassium clavulanate ODF in the library and average spectra of tablets of amoxicillin and potassium clavulanate were calculated from 4,200 to 10,000 $cm^{-1}$. The $r_T$ ranged from 34.42 to 99.69% with an average of 71.78%. The $r_T$ (Equation 1) of the two spectra $y_1$ (k) and $y_2$ (k) was calculated as the ratio of their covariance to the product of the two standard deviations $\sigma_{y1}$ and $\sigma_{y2}$. The value of $r_T$ ranges from −1 (inverted spectra) to +1 (identical spectra) and is expressed as a percentage.

$$r_T = \frac{Cov(y_1(k), y_2(k))}{\sigma_{y_1}\sigma_{y_2}} \quad (1)$$

## Construction of the NIR Quantitative Model

Calibration models were constructed using the PLS1 algorithm (PLS regression for one *y*-variable) (Brereton, 2000; Burns and Ciurczak, 2008) available in the Quant 2 package of OPUS v6.5 software. The Rank value is the number of main factors in building the PLS model. Validation methods of calibration model include a Test Set Validation (TSV) and Leave-One-Out Cross Validation (LOOCV). In the relevant Figures and Tables, rank is the number of PLS latent variables (LV), which is determined by a one-sided *F*-test on PRESS (Equation 2). $R^2$ (Equation 3) is the coefficient of determination, and gives the percentage of variance present in the true component values, which is reproduced in the prediction. M is the number of samples of the validation set. $Y_m$ is the mean of true concentration values. $Differ_i$ (Equation 4) is the difference between the true value and predicted value. RMSEP (Equation 5) is the root-mean-standard error of prediction in TSV. RMSECV (Equation 6) is the root-mean-standard error of LOOCV. Principal Component Analysis (PCA) scores indicate the position (coordinates) of the samples. PCA is calculated on the basis of calibration spectra.

$$PRESS = \sum_{i=1}^{M} \left(Differ_i\right)^2 \quad (2)$$

$$R^2 = \left(1 - \frac{\sum_{i=1}^{M} (Differ_i)^2}{\sum_{i=1}^{M} (Y_i - Y_m)^2}\right) \times 100 \quad (3)$$

$$Differ_i = Y_i^{true} - Y_i^{pred} \quad (4)$$

[1]Chinese Pharmacopeia 2015th Volume IV.103–104.

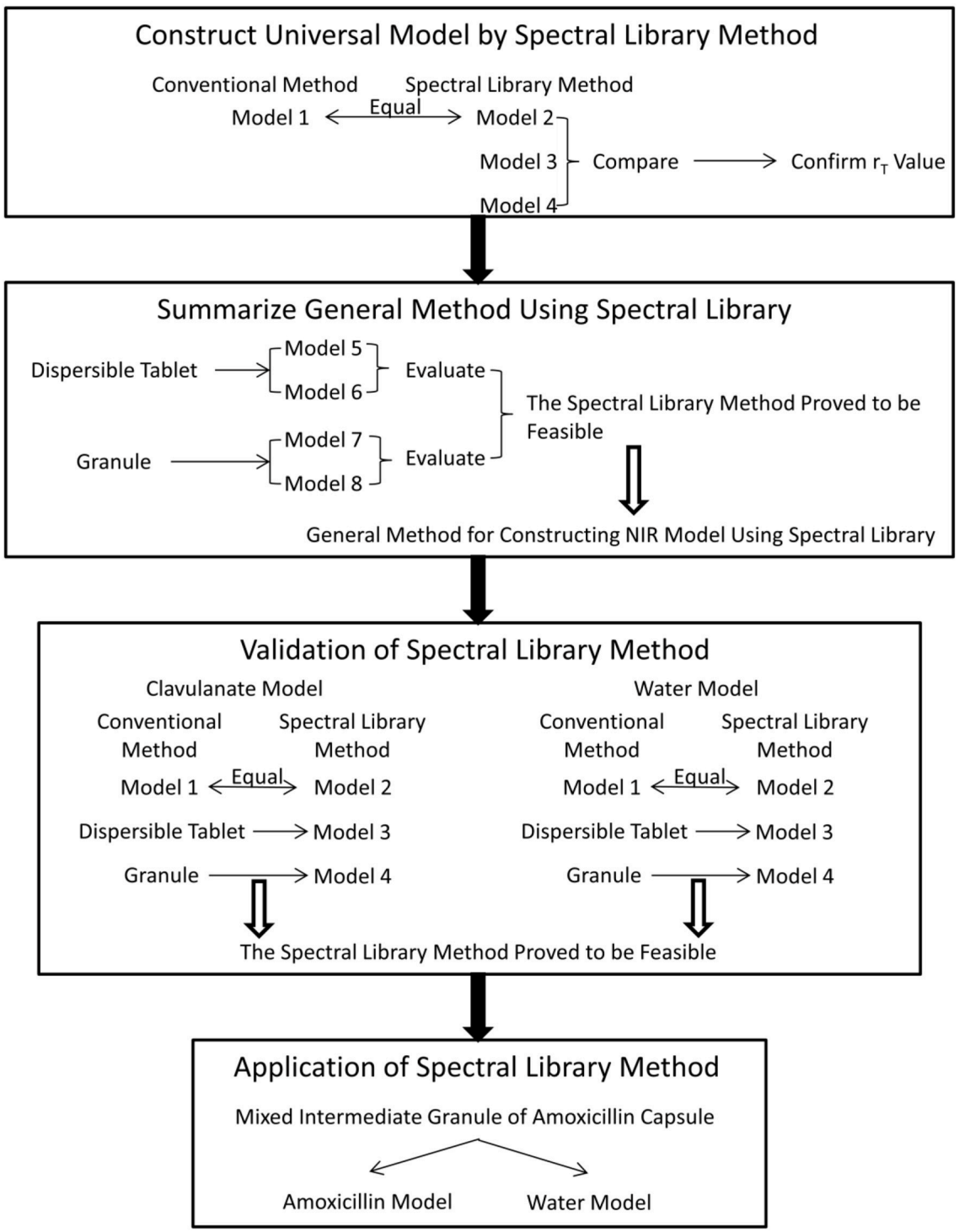



**FIGURE 1 |** Experimental design and list of all models.

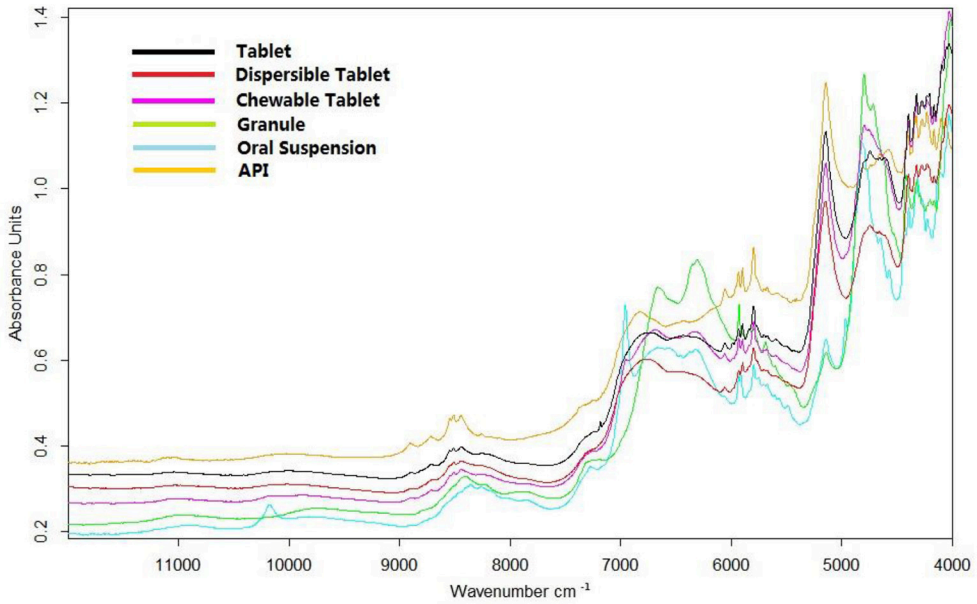



**FIGURE 2 |** Representative spectra of a tablet, dispersible tablet, chewable tablet, granule, and oral suspension of amoxicillin and potassium clavulanate, and the spectrum of amoxicillin.

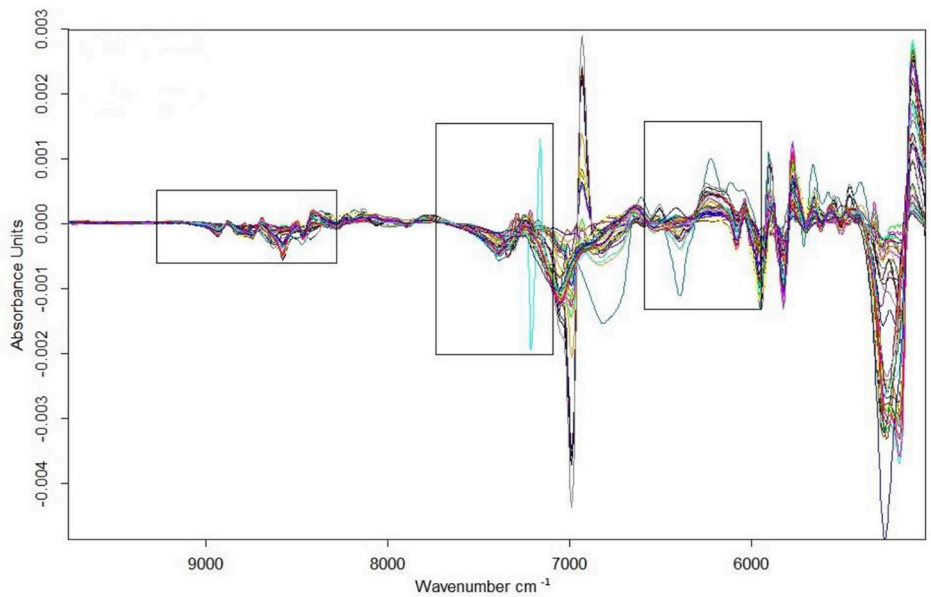



**FIGURE 3 |** Spectra of calibration samples of model 1 after being first-derivative preprocessed with a modeling spectral region.

$$\mathrm{RMSEP} = \sqrt{\frac{1}{M_t} \cdot \sum_{i=1}^{M_t} (\mathrm{Differ}_i)^2} \quad (5)$$

$$\mathrm{RMSECV} = \sqrt{\frac{1}{M_l} \cdot \sum_{i=1}^{M_l} (\mathrm{Differ}_i)^2} \quad (6)$$

## Conventional Method of Construction of a Universal Quantitative Model

A universal model was constructed based on our reported method (Chong et al., 2016). That is, all sample spectra were grouped into hierarchical clusters based on the Euclidean distance calculated from the Ward algorithm, and 19 groups were set according to the sample-selection strategy (Jia

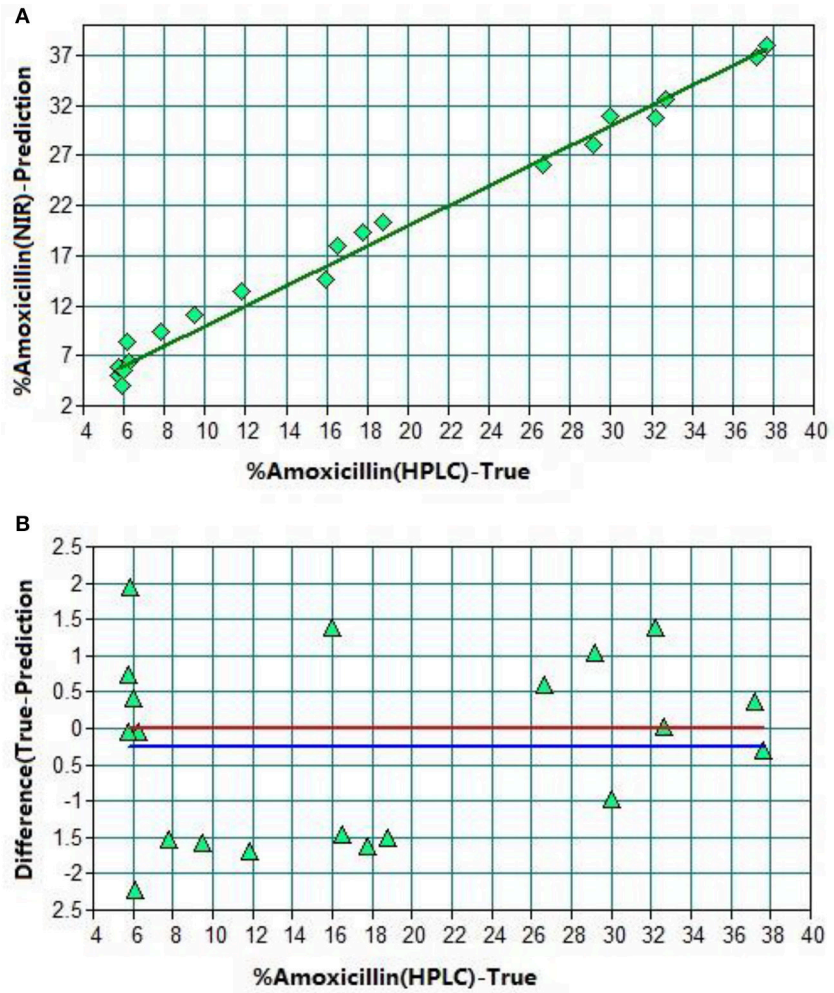



**FIGURE 4 |** TSV of model 1. **(A)** Plot of prediction values vs. true values. **(B)** Plot of difference vs. true values. The red line shows the zero line of difference. The blue line shows deviation of the zero line.

et al., 2011). Three random samples from each cluster were selected. Two of these samples were composed of the calibration set, and the remaining one was the validation set. Sixty-four spectra were selected to establish the NIR quantitative model to analyze the content of amoxicillin, potassium clavulanate, and water. Some test spectra for which the prediction differences were greater than the expected values, were transferred to the calibration set to optimize the model.

### Construction of a Universal Quantitative Model Using the NIR Spectral Library

All spectra were sequenced *via* the $r_T$ value. One spectrum was selected according to differences in $r_T$ values to construct a NIR quantitative model. Two-thirds of these spectra were composed of the calibration set; whereas one-third of the spectra were in the test set. Some spectra from the test set, for which the prediction difference was greater than expected, were transferred to the calibration set to optimize the model. These spectra could be used to analyze the content of amoxicillin, potassium clavulanate, and water.

### Validation of the Accuracy of the NIR Quantitative Model

The accuracy of NIR quantitative models was evaluated by Prediction Difference, which was the difference between the predicted content and reference content of amoxicillin, clavulanate, and water.

$$\text{Prediction Difference} = |\text{Prediction Content} - \text{Reference Content}|$$

## Sample Measurements

Here, 377 batches of amoxicillin and potassium clavulanate ODFs from post-marketing surveillance were measured in two time periods: 137 samples were measured for about 3 months in 2012, and the others were measured for about 6 months in 2014.

Spectra of 211 amoxicillin capsule samples were acquired for PAT control at ZhuHai United Laboratories (Guangdong Sheng, China) for about 7 months in 2016.

## EXPERIMENTAL DESIGN

Four steps were designed in the experiment (**Figure 1**). At first, a universal model (model 1) of amoxicillin for all amoxicillin and potassium clavulanate ODFs was constructed using a conventional calibration method for sample selection. Then, the NIR spectral library was used for modeling (model 2). Model 1 was used as a reference model to compare with model 2. If the results analyzed by model 2 were close to those analyzed by model 1, the spectral library was effective. Simultaneously, the appropriate difference between $r_T$ values of adjacent spectra in the calibration set was tested (models 2, 3, and 4). At the second step, models for a dispersible tablet (models 5 and 6) and models for a granule (models 7 and 8) were constructed using a general method for constructing a NIR model using spectral library. If the general models performed well, it was validated by constructing models for analyzing potassium clavulanate (clavulanate models 1, 2, 3, and 4) and water content (water models 1, 2, 3, and 4) in the third step. Finally, the spectral-library method was applied to a real PAT control. Models for analyzing amoxicillin and water content in mixed intermediate granules of amoxicillin capsules were constructed.

## RESULTS AND DISCUSSION

Representative spectra of a tablet, dispersible tablet, chewable tablet, granule, and oral suspension of amoxicillin and potassium clavulanate are shown in **Figure 2**. The spectra of a tablet, dispersible tablet, and chewable tablet are similar. Due to their prescription, low strength, and production process, the spectra of granule and oral suspension are quite different from those of a tablet, dispersible tablet, and chewable tablet. The spectrum of the amoxicillin API (**Figure 2**) was similar to the spectra of amoxicillin and potassium clavulanate ODFs in some spectral regions, such as the bands between 8,300 and 9,500 $cm^{-1}$ (overtone of C-H stretching vibrations), and between 5,300 and 6,500 $cm^{-1}$, 4,200 and 4,800 $cm^{-1}$ (overtone of C = O bonds). The calibration models analyzing amoxicillin could be set up on the basis of these spectral ranges.

### Universal Quantitative Model for Amoxicillin ODFs

The universal quantitative model for amoxicillin set up using the conventional method was called "amoxicillin model 1" (model 1). The spectral range employed for model 1 is shown in **Figure 3**. **Figure 4** shows the result of test-set validation of model 1.

After optimization, there were 44 sample spectra for the calibration set (training set) and 20 for the validation set (test set). $R^2$ was found to be 98.83% with RMSEP 1.23% (**Table 1**). The average predicted difference between the predicted content and reference content of amoxicillin was only 1.3%. Hence, this NIR method could be a replacement of the HPLC method.

The universal quantitative model for amoxicillin constructed using the NIR spectral library was called "amoxicillin model 2" (model 2). Model 2 and the subsequent models were optimized and validated by the same method as that used for model 1. The difference in $r_T$ values between adjacent spectra using model 2 was about 1.0%. The results of the two models were close (**Table 1**), so they had the same analysis capacity for amoxicillin.

The average of the difference between $r_T$ values of adjacent spectra in the calibration set of model 1 was about 1.5%. The influence of the difference between $r_T$ values of adjacent spectra was also investigated. "Amoxicillin model 3" (model 3) and "amoxicillin model" 4 (model 4) were constructed with a difference of 2.0 and 0.8%, respectively. The prediction differences of 377 samples analyzed by models 1, 2, 3, and 4 were compared (**Table 1**). We found that the prediction differences of

**TABLE 1 |** Parameters and prediction differences of models 1, 2, 3, and 4.

| Parameter | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | Training set | Test set | Training set | Test set | Training set | Test set | Training set | Test set |
| Number of samples | 44 | 20 | 46 | 19 | 25 | 8 | s61 | 25 |
| Content range (%, mg/mg) | 5.78–39.20 | 5.88–37.65 | 5.17–53.56 | 5.23–39.65 | 4.85–57.86 | 5.87–29.31 | 4.85–57.86 | 5.45–53.33 |
| Wavenumber range ($cm^{-1}$) | 9426.6–8273.4, 7702.5–7124.0, 6549.3–5970.7 | | 9426.6–7124.0, 6549.3–5396.0, 4825.2–4246.6 | | 7702.5–7124.0, 6549.3–5396.0, 4825.2–4246.6 | | 9426.6–7124.0, 6549.3–5396.0, 4825.2–4246.6 | |
| Pre-processing method | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | |
| Rank | 5 | | 5 | | 4 | | 4 | |
| $R^2$ (%) | 98.83 | | 99.13 | | 99.57 | | 99.00 | |
| RMSEP (%) | 1.23 | | 1.25 | | 0.509 | | 1.29 | |
| $r_T$ Difference (%) | 1.5 | | 1.0 | | 2.0 | | 0.75 | |
| Samples whose prediction difference is >5% | 3.7% (14/377) | | 3.7% (14/377) | | 7.4% (28/377) | | 5.0% (19/377) | |
| Samples whose prediction difference is <1% | 46.2% (174/377) | | 43.0% (162/377) | | 29.0% (109/377) | | 40.0% (151/377) | |
| Average prediction difference | 1.3% | | 1.6% | | 2.0% | | 1.7% | |

models 1, 2, and 4 were close and less than that of model 3—especially with samples, whose deviation was >5%. When there were large differences between $r_T$ values of adjacent spectra, the calibration samples decreased and became less representative. A difference of 1.5% was suitable for modeling.

## Construction of NIR Quantitative Models for Specific ODFs of Amoxicillin Using a Spectral Library

We used dispersible tablets of amoxicillin and potassium clavulanate as an example to establish a universal quantitative model for one dosage form (**Table 2**). Calibration sample spectra were selected according to the $r_T$ value from a spectral library comprising 377 spectra of amoxicillin and potassium clavulanate ODFs. At first, only the calibration spectra of dispersible tablets were selected from the spectral library. The 78 spectra of dispersible tablets were sequenced by $r_T$ value, and 30 spectra were chosen with a difference between $r_T$ values of adjacent spectra of 1.0–1.5% to construct "amoxicillin model" 5 (model 5).

"Amoxicillin model 6" (model 6) was established in a similar way. It means that the spectra of all dosage forms were selected for calibration. The average $r_T$ value of 78 dispersible tablets was nearly the median value of $r_T$ of the 30 calibration spectra, among which there were 12 spectra for a dispersible tablet. Comparing the PCA-score distribution space of models 5 and 6, the calibration samples of model 5 covered almost all of the distribution space of dispersible tablets; whereas the calibration samples of model 6 covered more space than model 5 (**Figure 5**). The prediction results of 78 batches of dispersible tablets by models 1, 2, 5, and 6 are shown in **Table 3**. The prediction differences seen in models 5 and 6 were lower than in the other two models. It is clearly indicated that it was feasible to construct NIR quantitative models of dispersible tablets of amoxicillin and potassium clavulanate using a spectral library.

The prescription and production process of tablets/dispersible tablets and granules/oral suspensions are quite different. As a result, the spectra of those dosage forms differed greatly (**Figure 2**). On this occasion, granules were taken as an example to validate the feasibility of constructing NIR quantitative models using a spectral library.

Models 7 and 8 were established similar to models 5 and 6. Thirty spectra were selected with a difference between $r_T$ values of adjacent spectra of 1.0–1.5%. Calibration spectra from model 7 were chosen from 96 spectra of granules. Samples from model 8 were from all dosage forms in the spectral library. Because the

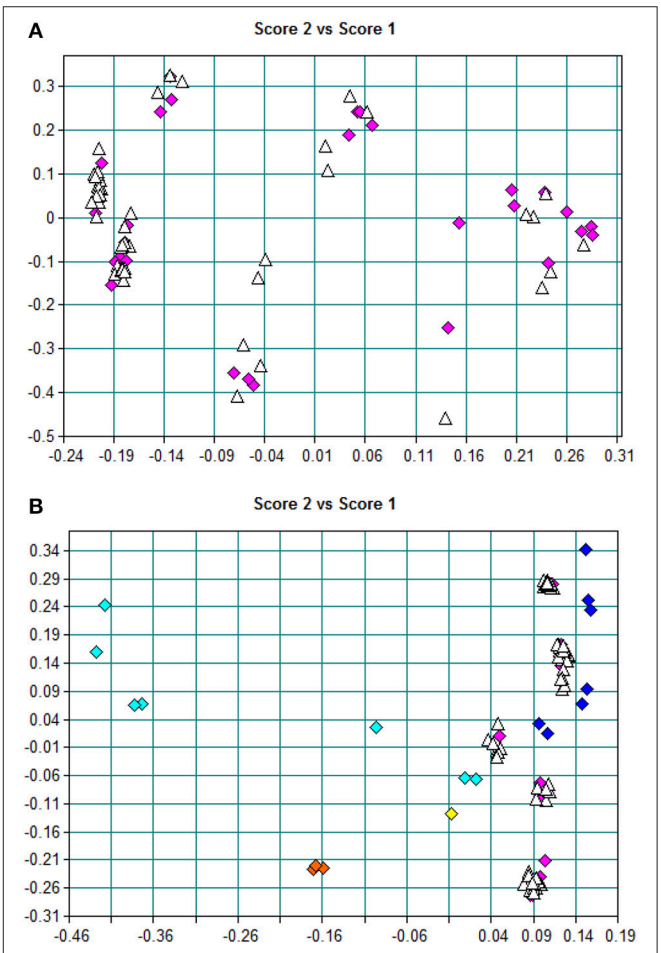



**FIGURE 5 | (A)** PCA scores of model 5; pink blocks represent training-set spectra, and white triangles represent other spectra of dispersible tablets in the spectral library. **(B)** PCA scores of model 6; blocks represent training-set spectra, dark-blue blocks represent tablets, pink blocks represent dispersible tablets, yellow blocks represent chewable tablets, light-blue blocks represent oral suspensions, orange blocks represent granules, and white triangles represent other spectra of dispersible tablets in the spectral library.

**TABLE 2 |** Parameters of models 5, 6, 7, and 8.

| Parameter | Model 5 | Model 6 | Model 7 | Model 8 |
|---|---|---|---|---|
| | Training set | Training set | Training set | Training set |
| Number of samples | 30 | 30 | 30 | 30 |
| Content range (%, mg/mg) | 18.57–40.32 | 7.97–57.86 | 4.85–8.05 | 4.77–22.27 |
| Wavenumber range ($cm^{-1}$) | 8277.2–7698.7, 5399.9–4821.2 | 6549.3–5970.7, 4825.2–4246.6 | 5797.2–5276.5 | 8851.9–8273.4, 6549.3–5970.7 |
| Pre-processing method | 1st derivative + vector normalization | 1st derivative + vector normalization | 1st derivative + vector normalization | 1st derivative + vector normalization |
| Rank | 4 | 4 | 5 | 5 |
| $R^2$ (%) | 97.48 | 96.46 | 96.28 | 94.07 |
| RMSECV (%) | 1.14 | 2.32 | 0.197 | 1.20 |

$r_T$ value of the oral suspension was close to that of a granule, 13 oral suspension spectra were comprised by model 8. A tablet spectrum was not included in model 8. The prediction values of all the granules included by the spectral library by models 1, 2, 7, and 8 are shown in **Table 3**. The results of models 7 and 8 were better.

Dispersible tablets of amoxicillin and potassium clavulanate could be analyzed equally well by models 5 and 6. Similar results could be obtained for granules by models 7 and 8. The $r_T$ value was critical for sample selection, but it was not necessary to choose the same dosage form as the samples to be measured.

A general method for constructing NIR quantitative models using a spectral library was summarized based on the experiments above (**Figure 6**). Firstly, the appropriate spectra of the samples to be measured were acquired, and the $r_T$ value calculated according to the definition of $r_T$ in the spectral library. Secondly, the calibration samples were selected based on the median $r_T$ value of samples to be measured. The difference between $r_T$ values of adjacent spectra in the calibration set was about 1.0–1.5%. The number of calibration samples should be $\geq$ 30, and their $r_T$ value should cover the range of samples to be measured. Finally, the model accuracy is validated by the samples to be measured. Appropriate sample spectra could be added to the calibration set to optimize the model if necessary.

## Validation of the General Method for Constructing a NIR Quantitative Model Using a Spectral Library

### Constructing NIR Quantitative Models for Potassium Clavulanate

A universal quantitative model for potassium clavulanate ("clavulanate model 1") was set up as shown in section Conventional Method of Construction of a Universal Quantitative Model. "Clavulanate model 2" was constructed by the general method as shown in **Figure 6**. The parameters and prediction difference for 377 samples in the spectral library of the two models (**Table 4**) indicated that the

**TABLE 3 |** Predictions of dispersible tablets and granules of amoxicillin and potassium clavulanate by amoxicillin models.

| Model | $r_T$ Difference | Dosage form | Number of samples | Samples whose prediction difference is >5% | Samples whose prediction difference is <1% | Average prediction difference |
|---|---|---|---|---|---|---|
| 1 | 1.5% | Dispersible tablets | 78 | 2.6% | 53.8% | 1.3% |
| | | Granules | 96 | 0% | 53.1% | 1.0% |
| 2 | 1.0% | Dispersible tablets | 78 | 6.4% | 43.6% | 1.9% |
| | | Granules | 96 | 0% | 51.0% | 1.2% |
| 5 | 0.7% | Dispersible tablets | 78 | 0% | 64.1% | 0.8% |
| 6 | 0.9% | Dispersible tablets | 78 | 0% | 52.6% | 1.2% |
| 7 | 1.6% | Granules | 96 | 0% | 100.0% | 0.1% |
| 8 | 1.1% | Granules | 96 | 0% | 81.2% | 0.8% |

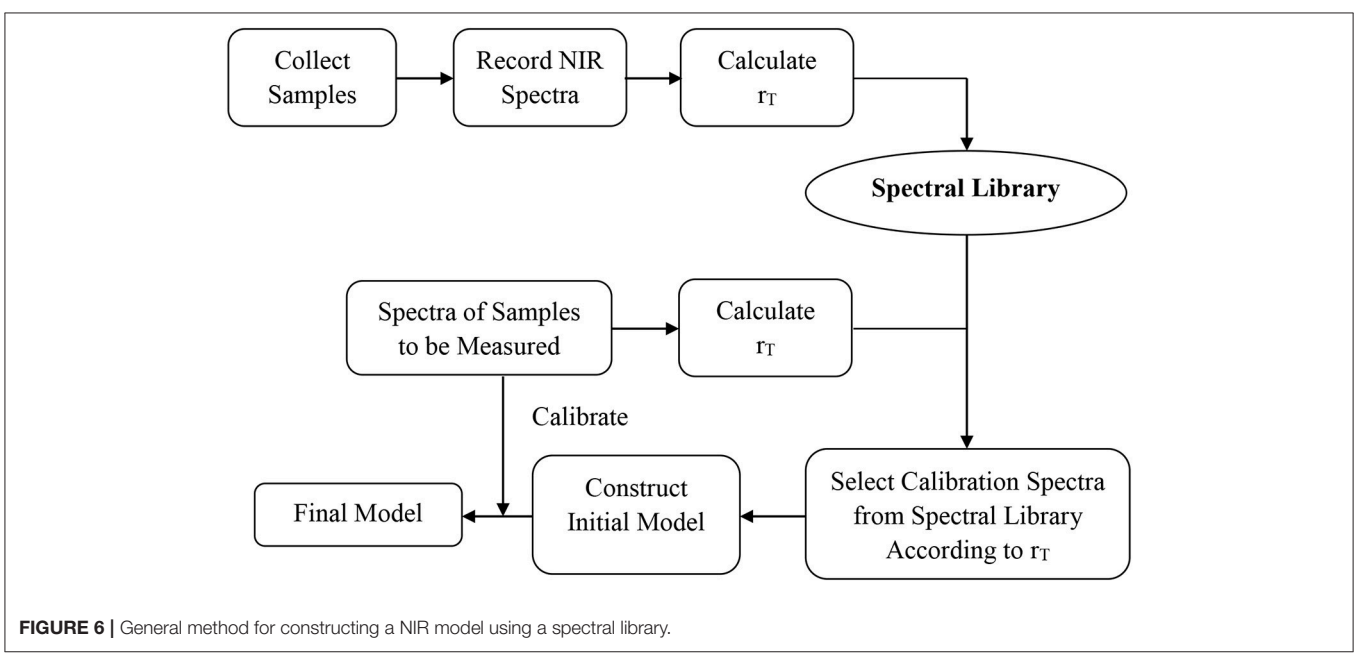



**FIGURE 6 |** General method for constructing a NIR model using a spectral library.

two methods of constructing models could lead to ideal results.

Similar to section Universal Quantitative Model for Amoxicillin ODFs, clavulanate model 3 (for dispersible tablets) and clavulanate model 4 (for granules) were constructed by the general method mentioned above. The prediction difference of 78 batches of dispersible tablets and 96 batches of granules by clavulanate models 3 and 4 were both <1.0% (**Table 4**). These two models were accurate and reliable. The feasibility of establishing a NIR quantitative model using the spectral library was also demonstrated.

### Constructing NIR Quantitative Models for Water Content

Universal quantitative models for water content (water model 1, 2, 3, 4) were established as shown in section Conventional Method of Construction of a Universal Quantitative Model, and the general method for a spectral library consequently resulted as shown in section Constructing NIR Quantitative Models for Potassium Clavulanate (**Table 5**). Water models 1 and 2 could be used to analyze all the dosage forms of amoxicillin and potassium clavulanate in the spectral library. Water models 3 and 4 could be used to analyze dispersible tablets and granules, respectively. **Table 5** shows that the prediction differences of the four models was <1.0%. These data further validated the validity of the general modeling method using a spectral library.

## Application of the Method for Constructing a NIR Quantitative Model Using a Spectral Library

The production process of amoxicillin capsules can be summarized as follows: granules are mixed with excipients after dry granulation of API and sieving; the mixed granules are then placed into capsules. The content of mixed intermediate granules of amoxicillin capsules ranged from 80.0 to 84.0%. The water content ranged from 12.1 to 13.0%. Mixed

**TABLE 4 |** Parameters and prediction differences of clavulanate models.

| Parameter | Clavulanate Model 1 | | Clavulanate Model 2 | | Clavulanate Model 3 | | Clavulanate Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | Training set | Test set | Training set | Test set | Training set | Test set | Training set | Test set |
| Number of samples | 53 | 16 | 46 | 19 | 30 | 12 | 32 | 10 |
| Content range (%, mg/mg) | 1.13–17.62 | 1.40–15.79 | 1.13–20.12 | 5.18–17.61 | 1.10–19.60 | 1.16–14.63 | 1.03–3.14 | 1.29–3.01 |
| Wavenumber range ($cm^{-1}$) | 7702.5–7124.0,6549.3–5970.7 | | 10001.3–8848.1, 6549.3–5396.0 | | 6549.3–5970.7,4825.2–4246.6 | | 10001.3–8848.1,6549.3–5396.0 | |
| Pre-processing method | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | |
| Rank | 5 | | 6 | | 6 | | 6 | |
| $R^2$ (%) | 99.63 | | 99.48 | | 99.63 | | 99.78 | |
| RMSEP (%) | 0.271 | | 0.386 | | 0.226 | | 0.0317 | |
| Samples whose prediction difference is >5% | 1.8% (7/377) | | 1.8% (7/377) | | 0% (0/78) | | 0% (0/96) | |
| Samples whose prediction difference is <1% | 69.0% (260/377) | | 72.1% (272/377) | | 60.0% (47/78) | | 93.7% (90/96) | |
| Average prediction difference | 0.8% | | 0.9% | | 0.8% | | 0.2% | |

**TABLE 5 |** Parameters and prediction differences of water models.

| Parameter | Water Model 1 | | Water Model 2 | | Water Model 3 | | Water Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | Training set | Test set | Training set | Test set | Training set | Test set | Training set | Test set |
| Number of samples | 47 | 19 | 46 | 16 | 30 | 10 | 32 | 11 |
| Content range (%, mg/mg) | 1.39–9.29 | 1.50–7.17 | 0.41–8.98 | 1.29–7.67 | 1.42–8.73 | 1.89–7.08 | 0.50–2.60 | 1.30–2.28 |
| Wavenumber range ($cm^{-1}$) | 9403.5–7498.1 | | 10502.8–9951.2, 7752.7–5546.4 | | 10502.8–8848.1, 8304.2–7197.3 | | 8851.9–7748.8, 6101.9–5546.4 | |
| Pre-processing method | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | | 1st derivative + vector normalization | |
| Rank | 3 | | 5 | | 6 | | 2 | |
| $R^2$ (%) | 99.33 | | 99.67 | | 99.23 | | 93.59 | |
| RMSEP (%) | 0.174 | | 0.117 | | 0.13 | | 0.0737 | |
| Samples whose prediction difference is >5% | 0% (0/377) | | 0% (0/377) | | 0% (0/78) | | 0% (0/96) | |
| Samples whose prediction difference is <1% | 87.0% (328/377) | | 80.9% (305/377) | | 47.4% (37/78) | | 96.9% (93/96) | |
| Average prediction difference | 0.4% | | 0.5% | | 1.0% | | 0.4% | |

**TABLE 6 |** Parameters and prediction difference of models for the amoxicillin capsule constructed using a spectral library.

| Parameter | Amoxicillin Model | | Water Model | |
|---|---|---|---|---|
| | Training set | Test set | Training set | Test set |
| Number of samples | 38 | 13 | 42 | 11 |
| Content range (%, mg/mg) | 7.97–84.40 | 8.05–84.31 | 1.42–12.80 | 1.89–12.60 |
| Wavenumber range ($cm^{-1}$) | 7702.5–7124, 6549.3–5970.7, 4825.2–4246.6 | | 9955.1–8848.1, 7752.7–7197.3 | |
| Preprocessing method | 1st derivative + vector normalization | | 1st derivative + vector normalization | |
| Rank | 5 | | 5 | |
| $R^2$ (%) | 99.91 | | 99.38 | |
| RMSEP (%) | 0.854 | | 0.256 | |
| Samples whose prediction difference is >5% | 0% (0/211) | | 0% (0/211) | |
| Samples whose prediction difference is <1% | 41.2% (87/211) | | 65.4% (138/211) | |
| Average prediction difference | 0.8% | | 0.4% | |

intermediate granules had only slight variability, so their NIR spectra were not suitable for a calibration set. We tried to set up NIR quantitative models for analyzing the content of amoxicillin and water in mixed intermediate granules of amoxicillin capsules using a spectral library of amoxicillin and potassium clavulanate ODFs because their spectra were similar.

The $r_T$ values of 211 samples were calculated according to the definition of $r_T$. The median value of $r_T$ of sample spectra was 91.33%. The maximum and minimum $r_T$ values were 99.29 and 88.98%, respectively. About 40 calibration spectra were selected from the spectral library. The difference in the adjacent spectra was 1.0–1.5%. The NIR model for amoxicillin was optimized by adding 16 spectra of mixed intermediate granules to the calibration set. The 13 spectra of mixed intermediate granules were added to the calibration set of the model for water content. Then, NIR quantitative models for the content of amoxicillin and water were constructed (**Table 6**). The prediction difference between the two models was small, so they could be used to analyze the content of amoxicillin and water of mixed granules rapidly during production.

## CONCLUSIONS

A NIR spectral library of amoxicillin and potassium clavulanate ODFs was established using a universal model. The similarity between NIR spectra was represented by the correlation coefficient $r_T$. About 30–50 calibration spectra were selected from the spectral library according to the median $r_T$ value to construct the NIR quantitative model. The difference in $r_T$ values between adjacent calibration spectra was about 1.0–1.5%. Compared with conventional modeling, this general method using a spectral library could be used to resolve sample-collection problems. This method requires calibration samples with an appropriate concentration range over a short time for PAT control. Furthermore, the quantitative models were more specific than models constructed by conventional methods. The proposed method offers a new and effective approach to solve the sample-selection problem in PAT modeling.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

## REFERENCES


Blanco, M., and Alcalá, M. (2006). Simultaneous quantitation of five active principles in a pharmaceutical preparation: development and validation of a near infrared spectroscopic method. *Eur. J. Pharm. Sci.* 27, 280–286. doi: 10.1016/j.ejps.2005.10.008

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., and Pezuela, C. (1997). Strategies for constructing the calibration set in the determination of active principles in pharmaceuticals by near infrared diffuse reflectance spectrometry. *Analyst* 122, 761–765. doi: 10.1039/a700630f

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., and Pou, N. (2001). Influence of the procedure used to prepare the calibration sample set on the performance of near infrared spectroscopy in quantitative pharmaceutical analysis. *Analyst* 126, 1129–1134. doi: 10.1039/b102090k

Blanco, M., Coello, J., Iturriaga, H., Maspoch, S., and Serrano, D. (1998). Near-infrared analytical control of pharmaceuticals. A single calibration model from mixed phase to coated tablets. *Analyst* 123, 2307–2312. doi: 10.1039/a805946b

Brereton, R. G. (2000). Introduction to multivariate calibration in analytical chemistry. *Analyst* 125, 2125–2154. doi: 10.1039/b003805i

Burns, D. A., and Ciurczak, E. W. (2008). *Handbook of Near-Infrared Analysis, 3rd Edn.* Boca Raton, FL: CRC Press, 198.

Chong, X. M., Zou, W. B., Yao, S. C., and Hu, C. Q. (2016). Rapid analysis of the quality of amoxicillin and clavulanate potassium tablets using diffuse reflectance near-infrared spectroscopy. *AAPS Pharm. Sci. Tech.* 18, 1311–1317. doi: 10.1208/s12249-016-0602-3

Corti, P., Ceramelli, G., Dreassi, E., and Mattii, S. (1999). Near infrared transmittance analysis for the assay of solid pharmaceutical dosage forms. *Analyst* 124, 755–758. doi: 10.1039/a809800j

Dong, Y., Li, J., and Zhong, X. (2016). High-throughput prediction of tablet weight and trimethoprim content of compound sulfamethoxazole tablets for controlling the uniformity of dosage units by NIR. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 159, 78–82. doi: 10.1016/j.saa.2016.01.030

Dreassi, E., Ceramelli, G., Corti, P., Perruccio, P. L., and Lonardi, S. (1996). Application of near-infrared reflectance spectrometry to the analytical control of pharmaceuticals: ranitidine hydrochloride tablet production. *Analyst* 121, 219–222. doi: 10.1039/an9962100219

Feng, Y. C., Zhang, X. B., and Hu, C. Q. (2010). Construction of an identification system for non-invasive analysis of macrolides tablets using near

infrared diffuse reflectance spectroscopy. *J. Pharm. Biomed. Anal.* 51, 12–17. doi: 10.1016/j.jpba.2009.07.018

Gottfries, J., Depui, H., Fransson, M., Jongeneelen, M., Josefson, M., Langkilde, F. W., et al. (1996). Vibrational spectrometry for the assessment of active substance in metoprolol tablet: a comparison between transmission and diffuse reflectance near-infrared spectrometry. *J. Pharm. Biomed. Anal.* 14, 1495–1503. doi: 10.1016/0731-7085(96)01800-6

Hertrampf, A., Muller, H., Menezes, J. C., and Herdling, T. (2015). A PAT-based qualification of pharmaceutical excipients produced by batch or continuous processing. *J. Pharm. Biomed. Anal.* 114, 208–215. doi: 10.1016/j.jpba.2015.05.012

Jamrógiewicz, M. (2012). Application of the near-infrared spectroscopy in the pharmaceutical technology. *J. Pharm. Biomed. Anal.* 66, 1–10. doi: 10.1016/j.jpba.2012.03.009

Jia, Y. H., Liu, X. P., Feng, YC., and Hu, C. Q. (2011). A training set selection strategy for a universal near infrared quantitative model. *AAPS Pharm. Sci. Tech*. 12, 738–745 doi: 10.1208/s12249-011-9638-6

Khan, A. N., Khar, R. K., and Ajayakumar, P. V. (2016). Diffuse reflectance near infrared-chemometric methods development and validation of amoxicillin capsule formulations. *J. Pharm. Bioallied Sci.* 8, 152–160. doi: 10.4103/0975-7406.175973

Mafalda, M. C., and Lopes, J. A. (2009). Quality control of pharmaceuticals with NIR: from lab to process line. *Vibr. Spectr*. 49, 204–210. doi: 10.1016/j.vibspec.2008.07.013

Merckle, P., and Kovar, K. A. (1998). Assay of effervescent tablets by near-infrared spectroscopy in transmittance and reflectance mode: acetylsalicylic acid in mono and combination formulations. *J. Pharm. Biomed. Anal.* 17, 365–374. doi: 10.1016/S0731-7085(97)00194-5

Moffat, A. C., Trafford, A. D., Jee, R. D., and Graham, P. (2000). Meeting of the international conference on hormonisation's guidelines on validation of analytical procedures: quantification as exemplified by a near-infrared reflectance assay of paracetamol intact tablets. *Analyst* 125, 1341–1351. doi: 10.1039/b002672g

Möltgen, C.-V., Puchert, T., Menezes, J. C., Lochmann, D., and Reich, G. (2012). A novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process. *Talanta* 92, 26–37. doi: 10.1016/j.talanta.2011.12.034

Sarraguça, P. R., Ribeiro, O., Santos, M. C., Silva, M. C., and Lopes, J. A. (2014). A PAT approach for the on-line monitoring of pharmaceutical co-crystals formation with near infrared spectroscopy. *Int. J. Pharm.* 471, 478–484. doi: 10.1016/j.ijpharm.2014.06.003

Silva, M. A., Ferreira, M. H., Braga, J. W., and Sena, M. M. (2012). Development and analytical validation of a multivariate calibration method for determination of amoxicillin in suspension formulations by near infrared spectroscopy. *Talanta* 89, 342–351. doi: 10.1016/j.talanta.2011.12.039

United States Food and Drug Administration (2004). *Guidance for Industry: PAT – a Framework for Innovative Pharmaceutical Development*. Manufacturing, and quality assurance. Pharmaceutical CGMPs.

United States Pharmacopeial Convention (2015). *In-process Revision: <856> Near-Infrared Spectroscopy [EB/OL]. [2015-02-27]*. Available online at: www.usppf.com/pf/pub/data/v411/CHAIPR411c856.xml

Velagaleti, R., Burns, P. K., Gill, M., and Prothro, J. (2002). Impact of current good manufacturing practices and emission regulations and guidance on the discharge of pharmaceutical chemicals into the environment from manufacturing, use, and disposal. *Environ. Health Perspect.* 110, 213–220. doi: 10.1289/ehp.02110213

Wahl, P. R., Fruhmann, G., Sacher, S., Straka, G., Sowinski, S., and Khinast, J. G. (2014). PAT for tableting: inline monitoring of API and excipients via NIR spectroscopy. *Eur. J. Pharm. Sci.* 87, 271–278. doi: 10.1016/j.ejpb.2014.03.021

Zou, W. B., Feng, Y. C., Dong, J. X., Song, D. Q., and Hu, C. Q. (2013). A new strategy to iteratively update scalable universal quantitative models for the testing of azithromycin by near infrared spectroscopy. *Sci. China Chem.* 56, 533–540. doi: 10.1007/s11426-012-4807-3

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 05 June 2018
doi: 10.3389/fchem.2018.00202

# Low-Cytotoxicity Fluorescent Probes Based on Anthracene Derivatives for Hydrogen Sulfide Detection

*Xuefang Shang[1]*, Jie Li[1], Yaqian Feng[2], Hongli Chen[3], Wei Guo[1], Jinlian Zhang[2], Tianyun Wang[4] and Xiufang Xu[5]*

[1] *Key Laboratory of Medical Molecular Probes, School of Basic Medical Sciences, Xinxiang Medical University, Xinxiang, China,* [2] *School of Pharmacy, Xinxiang Medical University, Xinxiang, China,* [3] *School of Life Sciences and Technology, Xinxiang Medical University, Xinxiang, China,* [4] *Department of Biochemistry, Xinxiang Medical University, Xinxiang, China,* [5] *Department of Chemistry, Nankai University, Tianjin, China*

Owing to the role of $H_2S$ in various biochemical processes and diseases, its accurate detection is a major research goal. Three artificial fluorescent probes based on 9-anthracenecarboxaldehyde derivatives were designed and synthesized. Their anion binding capacity was assessed by UV-Vis titration, fluorescence spectroscopy, HRMS, $^1$HNMR titration, and theoretical investigations. Although the anion-binding ability of compound **1** was insignificant, two compounds **2** and **3**, containing benzene rings, were highly sensitive fluorescent probes for $HS^-$ among the various anions studied ($HS^-$, $F^-$, $Cl^-$, $Br^-$, $I^-$, $AcO^-$, $H_2PO_4^-$, $SO_3^{2-}$, Cys, GSH, and Hcy). This may be explained by the nucleophilic reaction between $HS^-$ and the electron-poor C=C double bond. Due to the presence of a nitro group, compound **3**, with a nitrobenzene ring, showed stronger anion binding ability than that of compound **2**. In addition, compound **1** had a proliferative effect on cells, and compounds **2** and **3** showed low cytotoxicity against MCF-7 cells in the concentration range of 0–150 $\mu g \cdot mL^{-1}$. Thus, compounds **2** and **3** can be used as biosensors for the detection of $H_2S$ *in vivo* and may be valuable for future applications.

**Keywords: fluorescent probe, hydrogen sulfide, 9-anthracenecarboxaldehyde, nucleophilic substitution, cytotoxicity**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Lingxin Chen,*
*Yantai Institute of Coastal Zone Research (CAS), China*
*Aldo Arrais,*
*Università degli Studi del Piemonte Orientale, Italy*

****Correspondence:***
*Xuefang Shang*
*xuefangshang@126.com*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *22 January 2018*
***Accepted:*** *15 May 2018*
***Published:*** *05 June 2018*

***Citation:***
*Shang X, Li J, Feng Y, Chen H, Guo W, Zhang J, Wang T and Xu X (2018) Low-Cytotoxicity Fluorescent Probes Based on Anthracene Derivatives for Hydrogen Sulfide Detection. Front. Chem. 6:202. doi: 10.3389/fchem.2018.00202*

## INTRODUCTION

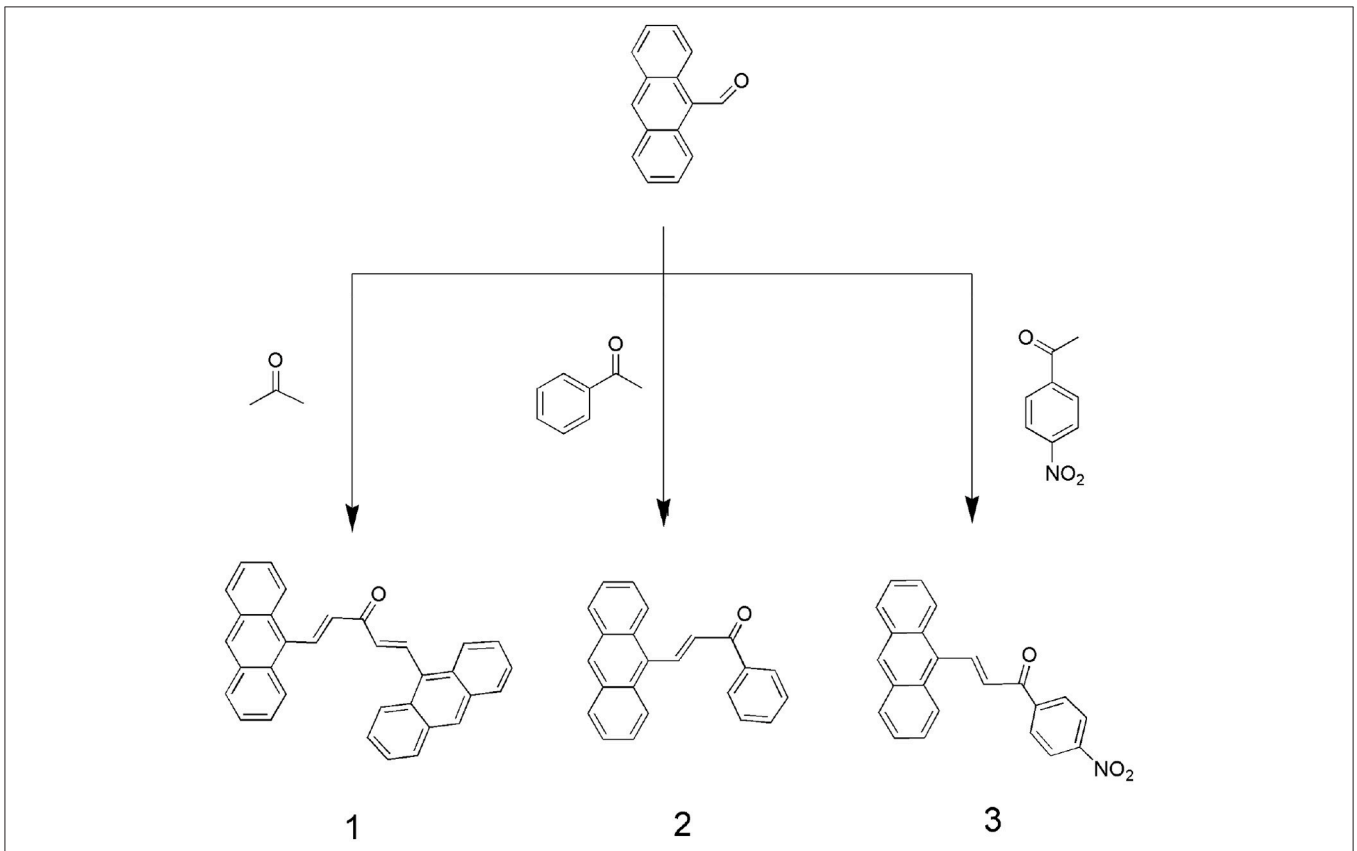
Hydrogen sulfide ($H_2S$) is a toxic gas with smell resembling rotten eggs. It is a bioactive gaseous signaling molecule, along with nitrous oxide (NO) and carbon monoxide (CO) (Kimura et al., 2012; Lisjak et al., 2013; Kimura, 2015; Mishanina et al., 2015). CO and NO are reactive oxygen species, whereas $H_2S$ gas is a scavenger of reactive oxygen species. Under certain pressure conditions, $H_2S$ can modulate mitochondria in mammalian cells. It also participates in many biochemical processes such as inflammation, blood pressure control, neuro-transmission, and ischemia reperfusion (Fu et al., 2012; Andreadou et al., 2015; Li F. et al., 2015; Wallace et al., 2015). $H_2S$ is also a relaxing agent that can act on smooth muscle and can serve as a modulator of cardiac function in cardiovascular therapy (Polhemus and Lefer, 2014; Barr et al., 2015; Chai et al., 2015; Holwerda et al., 2015). In addition, abnormal levels of $H_2S$ are associated with many diseases, oxygen sensing, and even death (Olson et al., 2006; Pandey et al., 2012). Therefore, the construction of fluorescent probe to detect $H_2S$ has important practical applications.

Traditional methods for determining the concentration of $H_2S$ in biological samples include colorimetric, electrochemical, chromatographic, metal-induced vulcanization, and fluorescence analyses (Tangerman, 2009; Shen et al., 2011). Fluorescent molecular probes are commonly used for detection tool in various fields, including in biological samples owing to their ability to convert chemical information into light signals with high sensitivity and selectivity. Hence, the development of fluorescent probes for the detection of $H_2S$ has attracted substantial research attention (Jiménez et al., 2003; Choi et al., 2009; Yu et al., 2012, 2014).

However, a few reports have focused on the development of fluorescent probes based on the binuclear character of $H_2S$ (Asthana et al., 2016; Das et al., 2016). Therefore, we used this approach to synthesize highly selective and sensitive fluorescent probes that can detect $H_2S$. Under physiological conditions, hydrogen sulfides exist as 30% $H_2S$ in a non-resolving state and 70% residual $HS^-$. Thus, $HS^-$ detection can serve as a proxy for $H_2S$. In this study, we designed and synthesized novel anthracene derivatives in which a -C=C- bond served as an interaction site (**Scheme 1**). The abilities of these compounds to bind to various anions ($HS^-$, $(n\text{-}C_4H_9)_4NF$ ($F^-$), $(n\text{-}C_4H_9)_4NCl$ ($Cl^-$), $(n\text{-}C_4H_9)_4NBr$ ($Br^-$), $(n\text{-}C_4H_9)_4NI$ ($I^-$), $(n\text{-}C_4H_9)_4NAcO$ ($AcO^-$), $(n\text{-}C_4H_9)_4NH_2PO_4$ ($H_2PO_4^-$), $Na_2SO_3$ ($SO_3^{2-}$), cysteine (Cys), glutathione(GSH), and homocysteine (Hcy) were assessed through UV-Vis titration, fluorescence spectroscopy, HRMS and $^1$HNMR titration for $HS^-$ sensitivity and selectivity. These compounds were also investigated for cytotoxicity to MCF-7 cells.

## MATERIALS AND METHODS

Most of the starting materials were obtained commercially. All reagents and solvents were of analytical grade. Sodium hydrosulfide, all anions, in the form of tetrabutylammonium salts such as $(n\text{-}C_4H_9)_4NF$, $(n\text{-}C_4H_9)_4NCl$, $(n\text{-}C_4H_9)_4NBr$, $(n\text{-}C_4H_9)_4NI$, $(n\text{-}C_4H_9)_4NAcO$, and $(n\text{-}C_4H_9)_4NH_2PO_4$, and amino acids (Cys, GSH, and Hcy) were purchased from Aladdin (Shanghai, People's Republic of China), stored in a vacuum desiccator containing self-indicating silica, and used without further purification. Tetrabutylammonium salts were dried for 24 h under a vacuum with $P_2O_5$ at 333 K before use. Dimethyl sulfoxide was distilled in vacuo after being dried with $CaH_2$. $^1$H NMR spectra were recorded using a Varian Unity Plus 400 MHz spectrometer. ESI-HRMS was performed using a Mariner apparatus. UV-Vis spectroscopy titration was performed using a Shimadzu UV2550 spectrophotometer at 289 K. Fluorometric titration was performed using an Eclipse fluorescence spectrophotometer (Agilent, Santa Clara, CA, USA) at 298 K. IR spectroscopy was performed using an IRTracer-100




**SCHEME 1 |** Synthesis routes of compounds **1**, **2**, and **3**.

instrument. The binding constants ($K_s$) were obtained by the non-linear least-squares method for data fitting.

Cells in logarithmic growth phase were seeded in 96-well plates at a density of $2.0 \times 10^4$ cells per well and cultured for 24 h. The culture medium was then replaced with 200 μL of Roswell Park Memorial Institute (RPMI) 1640 medium containing various concentrations of the compound, and the cells were further incubated for 24 h. Next, the cells were washed with phosphate buffered saline (PBS) three times, and 100 μL of culture medium and 20 μL of MTT solution were added to each well. After further incubation (4 h), the absorbance of each well was detected at 490 nm using a microplate reader (Thermo Multiskan MK3, Thermo Fisher Scientific, MA, USA). Plain cell culture medium was used as the control.

Compound **1** was synthesized according to previous methods (Ding et al., 2013). 9-Anthracenecarboxaldehyde (82.4 mg, 0.4 mmol) and acetone (35 mg, 0.6 mmol) were dissolved in ethanol (50 mL). Then, under stirring, an aqueous sodium hydroxide solution (2 mL, 0.04 mol·$L^{-1}$) was slowly added to the reaction flask. The mixture was stirred at room temperature for 6 h and adjusted to pH 5–6 with dilute hydrochloric acid (0.1 mol·$L^{-1}$) until the reaction was complete. The reaction was monitored by thin-layer chromatography. Typically, a precipitate formed and was collected by filtration. The solid was washed with high purity water and ethanol, and dried under a vacuum. Yield: 87%. $^1$H-NMR (400 MHz, $CDCl_3$, 298 K) δ 8.84 (d, J = 16.2 Hz, 1H), 8.52 (s, 1H), 8.38 (d, J = 8.3 Hz, 2H), 8.07 (d, J = 7.9 Hz, 2H), 7.69–7.47 (m, J = 88 Hz, 4H). $^{13}$C NMR (101 MHz, $CDCl_3$) δ 194.10, δ 147.53, δ 141.15, δ 135.40, δ 134.28, δ 129.71, δ 128.98, δ 128.60, δ 126.54, δ125.35. IR spectrum, ν $cm^{-1}$: 1668 (C=O); 1628 (C=C); 1593 (Ar-C=C); 999 (C=C-H). ESI-HRMS (*m/z*): 457.2 $(M + Na)^+$.

Compound **2** and **3** were synthesized according to the above procedure.

Compound **2**: $^1$H NMR (400 MHz, $CDCl_3$, 298 K) δ 8.83 (d, J = 15.8 Hz, 1H), 8.52 (s, 1H), 8.40–8.27 (m, J = 52 Hz, 2H), 8.18–8.00 (m, J = 72 Hz, 4H), 7.68–7.60 (m, 2H), 7.60–7.48 (m, 6H). $^{13}$C NMR (101 MHz, DMSO) δ 191.24,δ 140.88,δ 139.87,δ 137.75,δ 131.15,δ 129.15,δ 128.52,δ 127.32,δ 126.53,δ 125.50. IR spectrum, ν $cm^{-1}$: 3050 (Ar C-H); 1730 (C=O); 1560 (C=C); 720 (C=C-H). ESI-HRMS (*m/z*): 309.1 $(M + H)^+$, 331.1 $(M + Na)^+$.

Compound **3**: $^1$H NMR (400 MHz, $CDCl_3$, 298 K) δ 8.92 (d, J = 15.8 Hz, 1H), 8.92 (d, J = 15.8 Hz, 1H), 8.55 (s, 1H), 8.47 (s, 1H), 8.51–8.36 (m, J = 60.0 Hz, 3H), 8.42–8.22 (m, 6H), 8.28 (dd, J = 23.9 Hz, 8.3 Hz, 4H), 8.16–8.05 (m, J = 44 Hz, 2H), 8.15–8.04 (m, J = 44 Hz, 2H), 7.62–7.52 (m, J = 40 Hz, 4H), 7.65–7.52 (m, J = 52 Hz, 4H), 7.28 (s, 3H). $^{13}$C NMR (101 MHz, DMSO) δ 188.64,δ 150.36,δ 142.55,δ 131.59,δ 131.32,δ 129.47,δ 127.43,δ 126.15,δ 125.55,δ 124.41. IR spectrum, ν $cm^{-1}$: 1750 (C=O); 1590 (C=C); 1520 (N-O); 880 (C=N). ESI-HRMS (*m/z*): 376.1 $(M + Na)^+$.

# RESULTS AND DISCUSSION

## UV-Vis Spectral Titration

UV-Vis titration was performed in dimethyl sulfoxide by the stepwise addition of sodium hydrosulfide (**Figure 1**). For compound **1**, the presence of $HS^-$ resulted in an increase in the absorption intensity at 315 nm, but the spectral changes were very small. Furthermore, the addition of $F^-$, $Cl^-$, $Br^-$, $I^-$, $AcO^-$, $H_2PO_4^-$, $SO_3^{2-}$, Cys, GSH, or Hcy resulted in very weak spectral changes for compound **1**, and the binding capacity was negligible.

For compound **2**, the intensity of the absorption peak increased at 312 nm after the addition of sodium hydrosulfide. A hyperchromic effect was observed during the host-guest interaction process. The change in the UV-Vis spectrum was due to the interaction between sodium hydrosulfide and the electron-deficient C=C double bond (Zhao et al., 2012). However, the addition of $F^-$, $Cl^-$, $Br^-$, $I^-$, $AcO^-$, or $H_2PO_4^-$ did not cause a substantial spectral response for compound **2** (Figure S1), suggesting that the host-guest interaction was weak (Shao et al., 2009; Shang et al., 2013, 2015a). For compound **3**, the intensity of the absorption peak at 336 nm increased, and the absorption band was enhanced after $HS^-$ addition. However, the addition of $F^-$, $Cl^-$, $Br^-$, $I^-$, $AcO^-$, $H_2PO_4^-$, $SO_3^{2-}$, Cys, GSH, or Hcy resulted in a very weak spectral response, indicating that the host-guest interaction was negligible. These results suggested that compounds **2** and **3** both showed high sensitivity and selectivity for $HS^-$.

## Fluorescence Response

The photophysical responses of the three probes to various anions were examined. As shown in **Figure 2**, compound **1** showed an emission peak centered at 582 nm. After the addition of $HS^-$ to a solution of compound **1**, the spectral response of compound **1** was very weak, indicating that the binding ability was negligible.

For compound **2**, emission peaks were centered at 382 and 404 nm. After the addition of $HS^-$, the fluorescence emission was significantly quenched. No significant spectral changes were observed after titration of $F^-$, $Cl^-$, $Br^-$, $I^-$, $H_2PO_4^-$, $AcO^-$, $SO_3^{2-}$, Cys, GSH, or Hcy, indicating that compound **2** had an insignificant binding capacity for these anions (Figure S2A).

For compound **3**, there was almost no fluorescence response. After the addition of $HS^-$, a new emission peak at approximately 420 nm appeared, which was gradually accompanied by two shoulders centered at 402 and 440 nm. This fluorescence enhancement may be resulted from two possible signal transduction mechanisms: the inhibition of photo-electron transfer and binding induced by the guest's host molecules (Watanabe et al., 1998; Lee et al., 2002; Lin et al., 2006). However, no significant spectral changes were observed when compound **3** was titrated with $F^-$, $Cl^-$, $Br^-$, $I^-$, $H_2PO_4^-$, $AcO^-$, $SO_3^{2-}$, Cys, GSH, or Hcy, indicating that compound **3** did not significantly bind to these anions (Figure S2B). The fluorescence calibration curve for compound **3** after the addition of $HS^-$ indicated that the emission intensity was non-linear when various quantities of $HS^-$ were added to a solution with a certain concentration of compound **3** (Shang et al., 2012a).

## Binding Constant

The spectral responses of compound **1** after the addition of anions were very weak; hence, the binding constant could not be calculated. The UV-Vis spectral changes for compounds **2** and **3** were ascribed to the formation of host-guest (1:2)

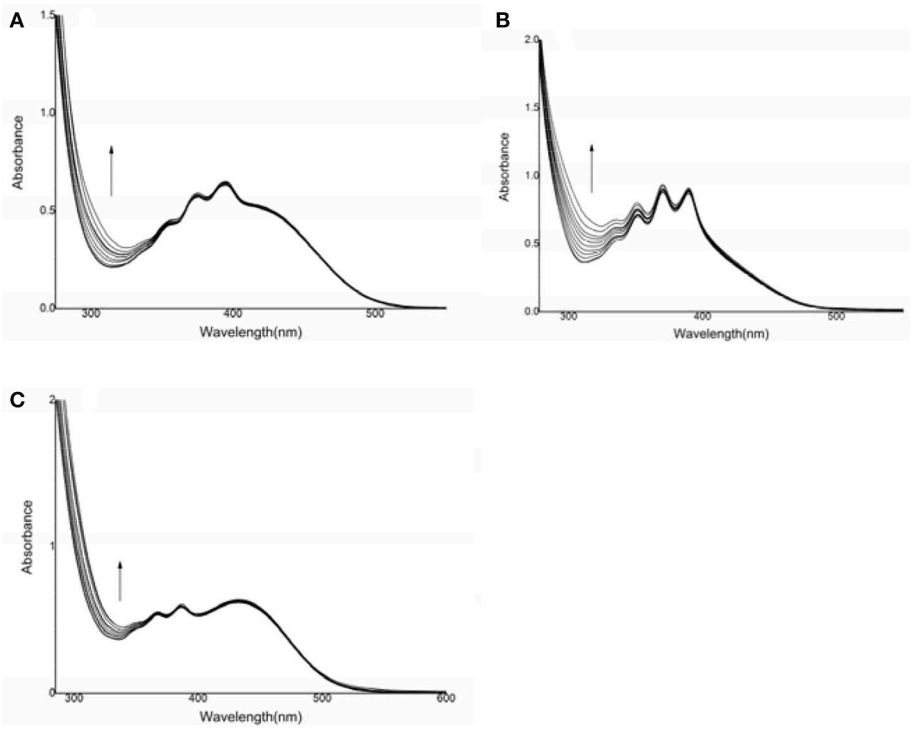



**FIGURE 1 |** UV-vis spectral changes of compounds **1**, **2,** and **3** after the addition of $HS^-$. **(A)** compound **1**: 6.90 × $10^{-5}$ mol·$L^{-1}$, $HS^-$: (0–76) × $10^{-5}$ mol·$L^{-1}$; **(B)** compound **2**: 1.46 × $10^{-4}$ mol·$L^{-1}$, $HS^-$: (0–2) × $10^{-3}$ mol·$L^{-1}$; **(C)** compound **3**: 1.1 × $10^{-4}$ mol·$L^{-1}$, $HS^-$: (0–16) × $10^{-4}$ mol·$L^{-1}$.

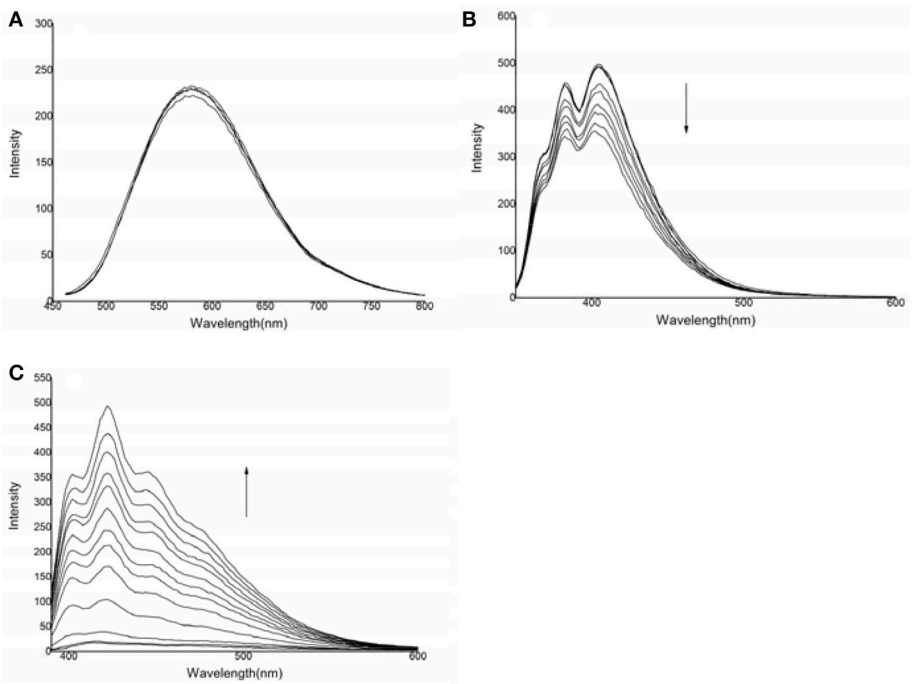



**FIGURE 2 |** Changes in the emission spectra of the three compounds in the presence of $HS^-$: **(A)** compound **1**: 6.9 × $10^{-5}$ mol·$L^{-1}$, $HS^-$: 0–20.7 × $10^{-5}$ mol·$L^{-1}$, $\lambda_{ex}$ = 442 nm; **(B)** compound **2**: 1.46 × $10^{-4}$ mol·$L^{-1}$, $HS^-$: 0–50.1 × $10^{-4}$ mol·$L^{-1}$, $\lambda_{ex}$ = 324 nm; **(C)** compound **3**: 1.1 × $10^{-4}$ mol·$L^{-1}$, $HS^-$: 0–7.7 × $10^{-4}$ mol·$L^{-1}$, $\lambda_{ex}$ = 368 nm.

complexes; when the absorbance intensity was greatest, the ratio of [H]/([H]+[G]) was approximately 0.3, according to a Job-plot (Figure S3). The binding constants were calculated by the non-linear least-squares method according to the UV-Vis data provided in **Table 1** (Bourson et al., 1993; Liu et al., 2001, 2004). It was shown that, the spectra changed little for compound **1**, and compounds **2** and **3** showed the strongest binding ability for $HS^-$ among the various anions tested. The anion binding abilities were in decreasing order: $HS^- >> SO_3^{2-} \sim Cys \sim GSH \sim Hcy \sim F^- \sim Cl^- \sim Br^- \sim I^- \sim AcO^- \sim H_2PO_4^-$. The standard deviations for the binding constants were $R_3 = 0.9941$ and $R_2 = 0.9945$. Among the three compounds, the standard deviation for compound **1** was not statistically significant, and those for compounds **2** and **3** were significant (compound **2**, **S** = 31.6011, compound **3**, **S** = 159.3298) (Figure S6). The anion binding ability could be attributed to the host-guest interactions and the match in space structures. It means that $HS^-$ ions strongly bound to these compounds, according to their binding constants (Shang et al., 2012b).

Compound **3** showed a stronger binding ability toward $HS^-$ ions than that of compound **2**, owing to the presence of a nitro group. The nitro group served as an electron-withdrawing group that enhanced the binding ability between the C=C double bond in compound **3** and $HS^-$. According to the HRMS data, the observed negative ion peak (418.0577) was the MS peak of the **3**-$HS^-$ complex (theoretical value: 418.0572) (Figure S4). In addition, there was no peak of $-CH_2-$ in the $^1$HNMR titration results, suggesting that the C=C double bond was broken during the interaction between compound **3** and $HS^-$ (Figure S5). Therefore, a possible host-guest binding mechanism was as follows. The first step was the Michael addition reaction of the conjugated system (Li J. et al., 2015). The first $HS^-$ ion was added to the C=C moiety as a nucleophile. Then, the second $HS^-$ ion attacked the active hydrogen atom (alpha-H) as an electrophile moiety, forming the final structure as shown in **Scheme 2**. The final structure was verified by mass spectrometry. The reaction of compound **3** with $HS^-$ was conducted in a simulated physiological environment, and the reaction product was subjected to a fluorescence analysis. A large increase in the fluorescence spectrum was observed.

**TABLE 1 |** Binding constants of the three compounds with various anions.

| Anion[a] | $K_S$ (1) | $K_S$ (2) | $K_S$ (3) |
|---|---|---|---|
| $HS^-$ | ND[b] | $(4.77 \pm 0.77) \times 10^5$ | $(1.07 \pm 0.45) \times 10^6$ |
| $F^-$, $Cl^-$, $Br^-$, $I^-$, $AcO^-$, $H_2PO_4^-$, $SO_3^{2-}$, Cys, GSH, Hcy | ND | ND | ND |

[a] *Anions was added in the form of sodium sulfide or tetra-n-butylammonium salts.*
[b] *The spectra changed little, and the binding constant could not be determined (ND).*

## Cytotoxicity Assessment

The cytotoxicity of the three compounds against MCF-7 cells was evaluated by MTT assays (Vibet et al., 2008; Jiang et al., 2014; Alemany et al., 2015; Jouvin et al., 2015; Moustakim et al., 2017) (**Figure 3**). Compound **1** had a proliferative effect on the cells, and compounds **2** and **3** in the range of 0–150 $\mu g \cdot mL^{-1}$ showed very low cytotoxicity. Cell viability was minimally affected (80% cell viability), when the concentrations of compounds **2** and **3** were increased to 150 $\mu g \cdot mL^{-1}$. In agreement with the determined binding constants, compounds **2** and **3** each showed a high binding capacity and low cytotoxicity and thus can be used to detect $HS^-$ *in vivo* (Gao et al., 2015; Shang et al., 2017). Compared with previous estimates in the literature (Zou et al., 2013; Lin et al., 2015), the cytotoxicity of the synthesized compounds was relatively low. Hence, these probes are favorable candidates for *in vitro* hydrogen sulfide detection.

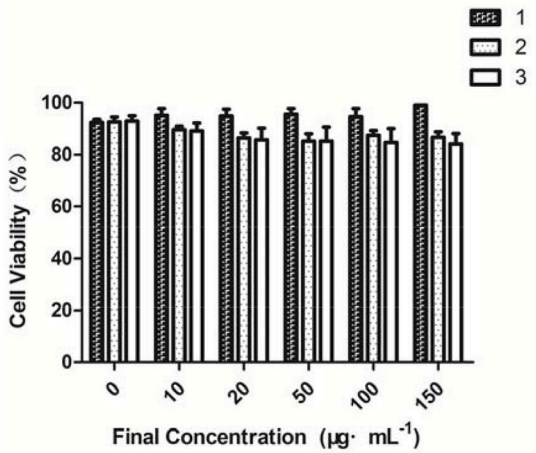
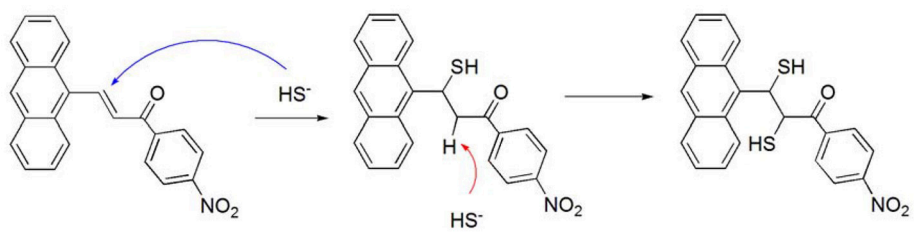



**FIGURE 3 |** Cell viability values (%), as estimated from MTT proliferation assays, vs. incubation concentrations of fluorescent probe.




**SCHEME 2 |** The possible interaction mechanism between compound **3** and $HS^-$.

## Theoretical Investigation

Among the three synthesized compounds, compound **3** showed the highest sensitivity and selectivity for $HS^-$ according to the binding constants. Consequently, the geometries were optimized for compound **3** and the combination product **3**-HS (**Figure 4**) based on the density functional theory method and the level of B3LYP/3-21G. The calculation was implemented in Gaussian03 (Frisch et al., 2003; Gao et al., 2017). As shown in **Figure 4**, the distance of the intramolecular hydrogen bond in compound **3** was 2.390 Å between the hydrogen atom of the interaction site (-HC=CH-) and the oxygen atom of the carbonyl group. According to previous studies (Ni et al., 2012; Maity et al., 2014), the existence of intramolecular hydrogen bonding and an electron-withdrawing group ($-NO_2$) increases the sensitivity. Hence, the stronger the electron-withdrawing effect is, the higher sensitivity for $HS^-$ this compound gets. The combination between compound **3** and $HS^-$ was also optimized. Our results indicated that the spatial structure of the host may change, as a result of the host-guest interaction. Therefore, the combination product (**3**-HS) existed in resonance form. The distance of the hydrogen bond (2.006 Å) indicated that a stable six-cycle was formed containing a sulfur atom and a hydrogen atom in a hydroxyl group (the resonance form of ketone) after compound **3** interacted with $HS^-$. These results also explained the strong ability of compound **3** to bind to $HS^-$.

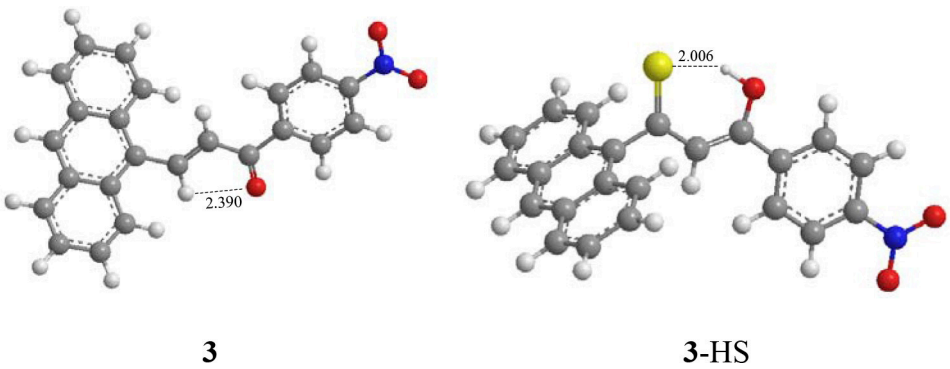



**FIGURE 4 |** Optimized geometries of compound **3** and the combination product **3**-HS.

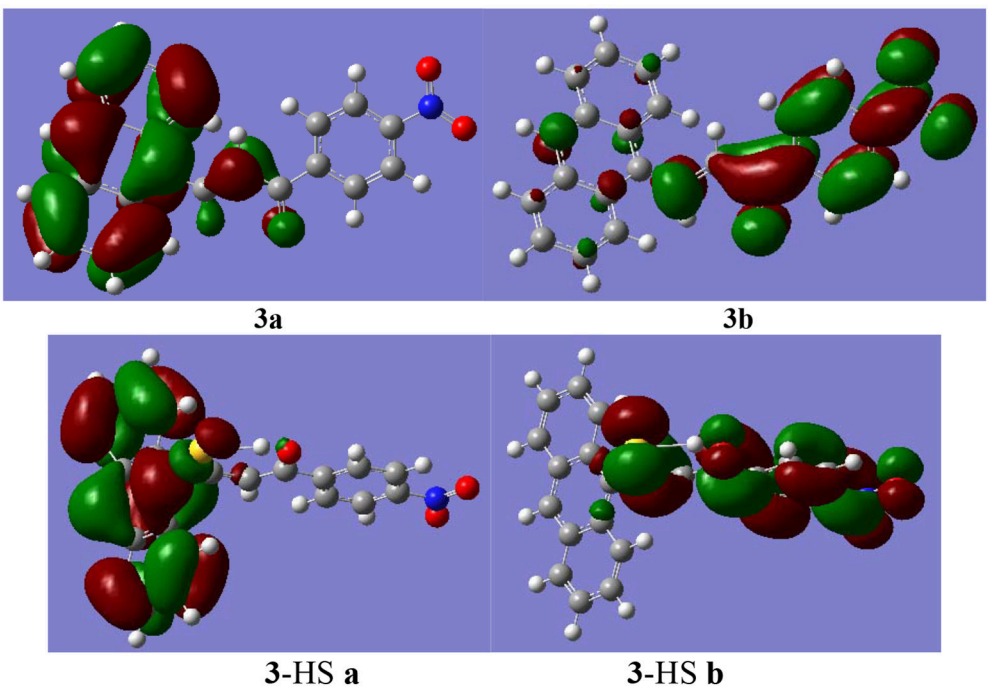



**FIGURE 5 |** The selected molecular frontier orbitals HOMO **(a)** and LUMO **(b)**.

In addition, the molecular frontier orbitals were introduced to explore the hyperchromic effect (by UV-Vis titration as described above). This effect was observed in the host-guest interaction process by the electron transition of the frontier orbital. The selected frontier orbitals for compound **3** and the host-guest complex are shown in **Figure 5**. An orbital analysis revealed that the highest occupied molecular orbital (HOMO) density in compound **3** was mainly localized on the anthracene moiety, whereas the lowest unoccupied molecular orbital (LUMO) density was localized on the nitrophenyl and ketone group moieties (Shang et al., 2015b). These results indicated that the electron transition of the highest HOMO resulted in a hyperchromic effect in the UV-Vis spectra.

## CONCLUSIONS

In conclusion, three compounds were synthesized, and their abilities to bind to various anions were detected by UV-Vis titration, fluorescence spectroscopy, HRMS, $^1$HNMR titration and theoretical investigations. Compounds **2** and **3** showed selectivity and sensitivity for $HS^-$. Notably, compound **3** showed the strongest sensing ability for $HS^-$ among the synthesized compounds. The mechanism underlying this interaction was the nucleophilic reaction between $HS^-$ and the electron-poor C=C double bond. Theoretical investigations also elucidated the role of molecular frontier orbitals in the hyperchromic effect. In addition, compounds **2** and **3** showed low cytotoxicity against MCF-7 cells in the concentration range of 0–150 $\mu g \cdot mL^{-1}$ and can be subsequently used as fluorescent probes to detect $H_2S$, $HS^-$, or $S^{2-}$ species *in vivo*. These results provide a probe with a novel sensing mechanism for hydrogen sulfide, based on the amphipolar character of the S atom of the new compounds to be used in practical applications to detect $H_2S$. Our finding establishes a basis for further applications of molecular probes.

## AUTHOR CONTRIBUTIONS

XS, and TW responsible for the experimental design. JL and YF responsible for the synthesis and properties of detection. WG and JZ responsible for the characterization of compounds. HC is responsible for the detection of cytotoxicity. XX is responsible for the quantitative calculation of the data.

## ACKNOWLEDGMENTS

This work was supported by funding from the Program for Science & Technology Innovation Talents in Universities of Henan Province (15HASTIT039), the Fluorescence Probe and Biomedical Detection Research Team of Xinxiang City (CXTD16001), the Xinxiang Medical University Graduate Scientific Research Innovation Support Project (YJSCX201638Y), and the Scientific and Technological Research Projects of Henan Province, China (172102210449, 182102311124). We would like to thank Editage (www.editage.com) and International Science Editing (http://www.internationalscienceediting.com) for English language editing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00202/full#supplementary-material

## REFERENCES

Alemany, L., Saunier, M., Alvarado-Cabrero, I., Quirós, B., Salmeron, J., Shin, H. R., et al. (2015). Human papillomavirus DNA prevalence and type distribution in anal carcinomas worldwide. *Int. J. Cancer* 136, 98–107. doi: 10.1002/ijc.28963

Andreadou, I., Iliodromitis, E. K., Szabo, C., and Papapetropoulos, A. (2015). Hydrogen sulfide and PKG in ischemia–reperfusion injury: sources, signaling, accelerators and brakes. *Basic Res. Cardiol.* 110:52. doi: 10.1007/s00395-015-0510-9

Asthana, S. K., Kumar, A., and Upadhyay, K. K. (2016). Efficient visualization of $H_2S$ via a fluorescent probe with three electrophilic centres. *Org. Biomol. Chem.* 14, 3690–3694. doi: 10.1039/C6OB00197A

Barr, L. A., Shimizu, Y., Lambert, J. P., Nicholson, C. K., and Calvert, J. W. (2015). Hydrogen sulfide attenuates high fat diet-induced cardiac dysfunction via the suppression of endoplasmic reticulum stress. *Nitric Oxide* 46, 145–156. doi: 10.1016/j.niox.2014.12.03

Bourson, J., Pouget, J., and Valeur, B. (1993). Ion-responsive fluorescent compounds. 4. Effect of cation binding on the photophysical properties of a coumarin linked to monoaza-and diaza-crown ethers. *J. Phys. Chem.* 97, 4552–4557. doi: 10.1021/j100119a050

Chai, Q., Lu, T., Wang, X. L., and Lee, H. C. (2015). Hydrogen sulfide impairs shear stress-induced vasodilation in mouse coronary arteries. *Pflügers Arch. Eur. J. Physiol.* 467, 329–340. doi: 10.1007/s00424-014-1526-y

Choi, M. G., Cha, S., Lee, H., Jeon, H. L., and Chang, S. K. (2009). Sulfide-selective chemosignaling by a $Cu^{2+}$ complex of dipicolylamine appended fluorescein. *Chem. Commun.* 7390–7392. doi: 10.1039/B916476F

Das, A. K., Goswami, S., Dutta, G., Maity, S., Kanti Mandal, T., Khanra, K., et al. (2016). A concentration dependent auto-relay-recognition by the same analyte: a dual fluorescence switch-on by hydrogen sulfide via Michael addition followed by reduction and staining for bio-activity. *Org. Biomol. Chem.* 14, 570–576.doi: 10.1039/C5OB02008E

Ding, J., Ge, Y., and Zhu, B. (2013). A highly selective fluorescent probe for quantitative detection of hydrogen sulfide. *Anal. Sci.* 29, 1171–1175. doi: 10.2116/analsci.29.1171

Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., et al. (2003). *Gaussian 03, revision A. 1*. Pittsburgh, PA: Gaussian Inc.

Fu, M., Zhang, W., Wu, L., Yang, G., Li, H., and Wang, R. (2012). Hydrogen sulfide ($H_2S$) metabolism in mitochondria and its regulatory role in energy production. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2943–2948. doi: 10.1073/pnas.1115634109

Gao, M., Yu, F., Chen, H., and Chen, L. (2015). Near-infrared fluorescent probe for imaging mitochondrial hydrogen polysulfides in living cells and *in vivo*. *Anal. Chem.* 87, 3631–3638. doi: 10.1021/ac5044237

Gao, M., Yu, F., Lv, C., Choo, J., and Chen, L. (2017). Fluorescent chemical probes for accurate tumor diagnosis and targeting therapy. *Chem. Soc. Rev.* 46, 2237–2271. doi: 10.1039/C6CS00908E

Holwerda, K. M., Karumanchi, S. A., and Lely, A. T. (2015). Hydrogen sulfide: role in vascular physiology and pathology. *Curr. Opin. Nephrol. Hypertens.* 24, 170–176. doi: 10.1097/MNH.0000000000000096

Jiang, P., Liu, Q., Liang, Y., Tian, J., Asiri, A. M., and Sun, X. (2014). A cost-effective 3D hydrogen evolution cathode with high catalytic activity: FeP nanowire

array as the active phase. *Angew. Chem. Int. Ed Engl.* 53, 12855–12859. doi: 10.1002/anie.201406848

Jiménez, D., Martínez-Máñez, R., Sancenón, F., Ros-Lis, J. V., Benito, A., and Soto, J. (2003). A new chromo-chemodosimeter selective for sulfide anion. *J. Am. Chem. Soc.* 125, 9000–9001. doi: 10.1021/ja0347336

Jouvin, K., Matheis, C., and Goossen, L. J. (2015). Synthesis of aryl tri- and difluoromethyl thioethers via a C-H-thiocyanation/fluoroalkylation cascade. *Chemistry* 21, 14324–14327. doi: 10.1002/chem.201502914

Kimura, H. (2015). Signaling molecules: hydrogen sulfide and polysulfide. *Antioxid. Redox Signal.* 22, 362–376. doi: 10.1089/ars.2014.5869

Kimura, H., Shibuya, N., and Kimura, Y. (2012). Hydrogen sulfide is a signaling molecule and a cytoprotectant. *Antioxid. Redox Signal.* 17, 45–57. doi: 10.1089/ars.2011.4345

Lisjak, M., Teklic, T., Wilson, I. D., Whiteman, M., and Hancock, J. T. (2013). Hydrogen sulfide: environmental factor or signalling molecule? *Plant Cell Environ.* 36, 1607–1616. doi: 10.1111/pce.12073

Lee, D. H., Im, J. H., Lee, J. H., and Hong, J. I. (2002). A new fluorescent fluoride chemosensor based on conformational restriction of a biaryl fluorophore. *Tetrahedron Lett.* 43, 9637–9640. doi: 10.1016/S0040-4039(02)02443-7

Li, F., Zhang, P., Zhang, M., Liang, L., Sun, X., Bao, A. et al. (2015). Effects of hydrogen sulfide on ozone-induced features of chronic obstructive pulmonary disease. *Eur. Respir. J.* 46(suppl59):PA4114. doi: 10.1183/13993003

Li, J., Yin, C., and Huo, F. (2015). Chromogenic and fluorogenic chemosensors for hydrogen sulfide: review of detection mechanisms since the year 2009. *RSC Adv.* 5, 2191–2206. doi: 10.1039/C4RA11870G

Lin, V. S., Chen, W., Xian, M., and Chang, C. J. (2015). Chemical probes for molecular imaging and detection of hydrogen sulfide and reactive sulfur species in biological systems. *Chem. Soc. Rev.* 44, 4596–4618. doi: 10.1039/C4CS00298A

Lin, Z. H., Zhao, Y. G., Duan, C. Y., Zhang, B. G., and Bai, Z. P. (2006). A highly selective chromo-and fluorogenic dual responding fluoride sensor: naked-eye detection of $F^-$ ion in natural water via a test paper. *Dalton Trans.* 3678–3684. doi: 10.1039/B601282E

Liu, Y., Han, B. H., and Zhang, H. Y. (2004). Spectroscopic studies on molecular recognition of modified cyclodextrins. *Curr. Org. Chem.* 8, 35–46. doi: 10.2174/1385272043486061

Liu, Y., You, C. C., and Zhang, H. Y. (2001). *Supramolecular Chemistry*. Tian Jin, Nankai University Publication.

Maity, D., Bhaumik, C., Mondal, D., and Baitalik, S. (2014). Photoinduced intramolecular energy transfer and anion sensing studies of isomeric Ru II Os II complexes derived from an asymmetric phenanthroline–terpyridine bridge. *Dalton Trans.* 43, 1829–1845. doi: 10.1039/C3DT52186A

Mishanina, T. V., Libiad, M., and Banerjee, R. (2015). Biogenesis of reactive sulfur species for signaling by hydrogen sulfide oxidation pathways. *Nat. Chem. Biol.* 11, 457–464. doi: 10.1038/nchembio.1834

Moustakim, M., Clark, P. G., Trulli, L., Fuentes de Arriba, A. L., Ehebauer, M. T., Chaikuad, A., et al. (2017). Discovery of a PCAF bromodomain chemical probe. *Angew. Chem. Int. Ed Engl.* 56, 827–831. doi: 10.1002/anie.201610816

Ni, X. L., Tahara, J., Rahman, S., Zeng, X., Hughes, D. L., Redshaw, C., et al. (2012). Ditopic Receptors based on lower-and upper-rim substituted hexahomotrioxacalix [3] arenes: cation-controlled hydrogen bonding of anion. *Chem. Asian J.* 7, 519–527.doi: 10.1002/asia.201100926

Olson, K. R., Dombkowski, R. A., Russell, M. J., Doellman, M. M., Head, S. K., Whitfield, N. L., et al. (2006). Hydrogen sulfide as an oxygen sensor/transducer in vertebrate hypoxic vasoconstriction and hypoxic vasodilation. *J. Exp. Biol.* 209, 4011–4023. doi: 10.1242/jeb.02480

Pandey, S. K., Kim, K. H., and Tang, K. T. (2012). A review of sensor-based methods for monitoring hydrogen sulfide. *TrAC Trends Anal. Chem.* 32, 87–99. doi: 10.1016/j.trac.2011.08.008

Polhemus, D. J., and Lefer, D. J. (2014). Emergence of hydrogen sulfide as an endogenous gaseous signaling molecule in cardiovascular disease. *Circ. Res.* 114, 730–737. doi: 10.1161/CIRCRESAHA

Shang, X., Hao, Y., Wang, Y., Han, J., Zhai, Y., Jia, S. et al. (2012a). Influence of different substituents on anion binding ability in aromatic hydroxyl group derivatives: experiment and theory. *Curr. Anal. Chem.* 8, 392–399. doi: 10.2174/157341112801264950

Shang, X., Li, J., Guo, K., Ti, T., Wang, T., and Zhang, J. (2017). Development and cytotoxicity of Schiff base derivative as a fluorescence probe for the detection of l-Arginine. *J. Mol. Struct.* 1134, 369–373. doi: 10.1016/j.molstruc.2016.12.105

Shang, X., Li, W., Wei, X., Zhang, H., Fu, Z., Zhang, J., et al. (2015a). Synthesis, Bioactivity, and the Anion-Binding Property of 2-Sulfydryl-1, 3, 4-thiodiazole Derivatives. *Heteroatom Chem.* 26, 142–149. doi: 10.1002/hc.21239

Shang, X., Li, X., Li, C., Wang, Y., Zhang, J., and Xu, X. (2012b). Spectral response to oxy-anions based on ferrocenylphalene. *Inorgan. Chim. Acta* 385, 128–134. doi: 10.1016/j.ica.2012.01.044

Shang, X., Luo, L., Ren, K., Wei, X., Feng, Y., Li, X., et al. (2015b). Synthesis and cytotoxicity of azo nano-materials as new biosensors for l-Arginine determination. *Mater. Sci. Eng. C* 51, 279–286. doi: 10.1016/j.msec.2015.03.005

Shang, X., Tian, S., Xi, N., Li, Y., Liu, Y., Yin, Z., et al. (2013). Colorimeric and fluorescence ON–OFF probe for acetate anion based on thiourea derivative: theory and experiment. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 103, 276–281. doi: 10.1016/j.saa.2012.11.027

Shao, J., Yu, X., Xu, X., Lin, H., Cai, Z., and Lin, H. (2009). Colorimetric and fluorescent sensing of biologically important fluoride in physiological pH condition based on a positive homotropic allosteric system. *Talanta* 79, 547–551. doi: 10.1016/j.talanta.2009.02.023

Shen, X., Pattillo, C. B., Pardue, S., Bir, S. C., Wang, R., and Kevil, C. G. (2011). Measurement of plasma hydrogen sulfide *in vivo* and *in vitro*. *Free Radic. Biol. Med.* 50, 1021–1031. doi: 10.1016/j.freeradbiomed.2011.01.025

Tangerman, A. (2009). Measurement and biological significance of the volatile sulfur compounds hydrogen sulfide, methanethiol and dimethyl sulfide in various biological matrices. *J. Chromatogr. B* 877, 3366–3377. doi: 10.1016/j.jchromb.2009.05.026

Vibet, S., Goupille, C., Bougnoux, P., Steghens, J. P., Goré, J., and Mahéo, K. (2008). Sensitization by docosahexaenoic acid (DHA) of breast cancer cells to anthracyclines through loss of glutathione peroxidase (GPx1) response. *Free Radic. Biol. Med.* 44, 1483–1491. doi: 10.1016/j.freeradbiomed.2008.01.009

Wallace, J. L., Blackler, R. W., Chan, M. V., Da Silva, G. J., Elsheikh, W., Flannigan, K. L., et al. (2015). Anti-inflammatory and cytoprotective actions of hydrogen sulfide: translation to therapeutics. *Antioxid. Redox Signal.* 22, 398–410. doi: 10.1089/ars.2014.5901

Watanabe, S., Onogawa, O., Komatsu, Y., and Yoshida, K. (1998). Luminescent metalloreceptor with a neutral bis (acylaminoimidazoline) binding site: optical sensing of anionic and neural phosphodiesters. *J. Am. Chem. Soc.* 120, 229–230. doi: 10.1021/ja973263a

Yu, F., Han, X., and Chen, L. (2014). Fluorescent probes for hydrogen sulfide detection and bioimaging. *Chem. Commun.* 50, 12234–12249. doi: 10.1039/C4CC03312D

Yu, F., Li, P., Song, P., Wang, B., Zhao, J., and Han, K. (2012). An ICT-based strategy to a colorimetric and ratiometric fluorescence probe for hydrogen sulfide in living cells. *Chem. Commun.* 48, 2852–2854. doi: 10.1039/C2CC17658K

Zhao, Y., Zhu, X., Kan, H., Wang, W., Zhu, B., Du, B., and Zhang, X. (2012). A highly selective colorimetric chemodosimeter for fast and quantitative detection of hydrogen sulfide. *Analyst* 137, 5576-5580. doi: 10.1039/C2AN36106J

Zou, Q., Fang, Y., Zhao, Y., Zhao, H., Wang, Y., Gu, Y., and Wu, F. (2013). Synthesis and *in vitro* photocytotoxicity of coumarin derivatives for one- and two-photon excited photodynamic therapy. *J. Med. Chem.* 56, 5288-5294. doi: 10.1021/jm400025g

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 19 June 2018
doi: 10.3389/fchem.2018.00228

# MicroNIR/Chemometrics Assessement of Occupational Exposure to Hydroxyurea

*Roberta Risoluti* and Stefano Materazzi*

*Department of Chemistry, Sapienza - University of Rome, Rome, Italy*

Portable Near Infrared spectroscopy (NIRs) coupled to chemometrics was investigated for the first time as a novel entirely on-site approach for occupational exposure monitoring in pharmaceutical field. Due to a significant increase in the number of patients receiving chemotherapy, the development of reliable, fast, and on-site analytical methods to assess the occupational exposure of workers in the manufacture of pharmaceutical products, has become more and more required. In this work, a fast, accurate, and sensitive detection of hydroxyurea, a cytotoxic antineoplastic agent commonly used in chemotherapy, was developed. Occupational exposure to antineoplastic agents was evaluated by collecting hydroxyurea on a membrane filter during routine drug manufacturing process. Spectra were acquired in the NIR region in reflectance mode by the means of a miniaturized NIR spectrometer coupled with chemometrics. This MicroNIR instrument is a very ultra-compact portable device with a particular geometry and optical resolution designed in such a manner that the reduction in size does not compromise the performances of the spectrometer. The developed method could detect up to 50 ng of hydroxyurea directly measured on the sampling filter membrane, irrespective of complexity and variability of the matrix; thus extending the applicability of miniaturized NIR instruments in pharmaceutical and biomedical analysis.

**Keywords: MicroNIR, chemometrics, hydroxyurea, occupational exposure, pharmaceutics**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Daniel Cozzolino,*
*Central Queensland University, Australia*
*Marçal Plans Pujolras,*
*Nestle Purina PetCare Company, United States*

****Correspondence:***
*Roberta Risoluti*
*roberta.risoluti@uniroma1.it*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *01 March 2018*
***Accepted:*** *31 May 2018*
***Published:*** *19 June 2018*

***Citation:***
*Risoluti R and Materazzi S (2018) MicroNIR/Chemometrics Assessement of Occupational Exposure to Hydroxyurea. Front. Chem. 6:228. doi: 10.3389/fchem.2018.00228*

## INTRODUCTION

Hydroxyurea (HU) or hydroxycarbamide, is a non-alkylating hydroxylated urea analog mainly recognized as antineoplastic and antiviral agent (Spivak and Hasselbalch, 2011). The cytotoxic and genotoxic potential efficacy of hydroxyurea makes this molecule one of the most performing agent commonly used in chemotherapy (Spivak and Hasselbalch, 2011; Karsy et al., 2016; Liew et al., 2016). In addition, HU is usually involved in the treatment of Sickle Cell Disease (SCD) (Davies and Gilmore, 2003; Heeney and Ware, 2008; Italia et al., 2009; Flanagana et al., 2010; Candrilli et al., 2011), psoriasis (Yarbro and Leavell, 1969), Philadelphia-chromosome negative myeloproliferative syndromes (MPs) (Yarbro and Leavell, 1969), some types of solid cancers (Karsy et al., 2016), and in the therapy of HIV infection (Lori et al., 1994).

An important issue when dealing with HU is related to its harmful potential (Millicovsky et al., 1981; Woo et al., 2005) especially in prolonged exposure conditions (Elchuri et al., 2015; Broto et al., 2017), as it inhibits class I ribonucleotide reductase, leading to replication fork stalling (Quattrone et al., 2013; Liew et al., 2016). Workers involved in the manufacture of drugs, may be exposed to HU during manufacturing, transport, and distribution. In addition, as the number of patients receiving chemotherapy has considerably increased, there is a growing concern about the development of

reliable, fast and accurate methods to assess the occupational exposure of workers during drug manufacturing process.

A number of analytical methods have been developed to quantify hydroxyurea in biological fluids, including spectrophotometric measurements by colorimetric techniques (Milks and Janes, 1956; Davidson and Winter, 1963; Bolton et al., 1965; Sivakumar et al., 2013; Legranda et al., 2017), electroanalytical determination (Naik et al., 2015), Nuclear Magnetic Resonance (NMR) (Main et al., 1987; Sorg et al., 2005; De Marco et al., 2011), High Performance Liquid Chromatography (HPLC) (Pujari et al., 1997; Iyamu et al., 1998; Manouilov et al., 1998), Gas Chromatography coupled to Mass Spectrometry (GC-MS) (James et al., 2006; Kettani et al., 2009; Garg et al., 2015), and Liquid chromatography—tandem mass spectrometry (LC-MS/MS) (Dalton et al., 2005; Usawanuwat et al., 2014; Marahatta et al., 2016; Hai et al., 2017). Despite the copious literature for HU detection, the assay of HU may be cumbersome due to its molecular dimension, reactivity and ability to chemical and enzymatic degradation (Iyamu et al., 1998; Marahatta and Ware, 2017).

The National Institute for Occupational and Safety Health (NIOSH) (Naumann et al., 1996) has proposed the exposure control limits (ECL) for HU not exceeding 0.01 mg/m$^3$, as a consequence of the potential toxicity. Conventional chromatographic techniques (Osytek et al., 2008) usually require an accurate sample clean-up to extract HU from a filter membrane and eliminate matrix interferences. All these procedures may be critical in estimating a tiny amount of HU and may lead to sample modification (Osytek et al., 2008). To overcome these problems, spectroscopic techniques have been largely proposed to give both qualitative and quantitative information about complex samples (Zontova et al., 2016; Materazzi et al., 2017a,b). In addition, multivariate statistical analysis has already proved to be helpful in interpreting complex spectral signals (Oliveri et al., 2011; Risoluti et al., 2016a,b, 2018; Materazzi et al., 2017c).

In this work, Near Infrared Spesctroscpy is proposed as a rapid and non-destructive technique to detect and quantify HU on a glass fiber filter in order to assess a novel procedure for occupational exposure estimation. A very ultra-compact portable instrument named MicroNIR (45-mm diameter, 42-mm height and 60-g operating weight) entirely powered (5 V) and controlled via USB port of a portable computer, was used to acquire spectra; and chemometrics tools were considered to perform real-time estimation of HU. A key feature of our portable MicroNIR/Chemometrics approach is mainly related to the possibility of directly analyze samples without any pre-treatment or extraction. In addition, the method is simple and time-saving, and it can achieve the same outcomes as the conventional spectrometer.

# MATERIALS AND METHODS

## Materials

Hydroxyurea reference standard was purchased as powder from Sigma-Aldrich. Glass fiber filters with 2.5-cm diameter, 1-μm pore size, and 790-μm thickness (Merk Millipore) were used as membrane to collect HU. Sampling was performed by the means of a Chronos sampling device (Zambelli Srl) operated at a flow rate of 3.5 L/min for 15 min, in order to mimic occupational exposure (not exceeding 3.5 μg/filter). Reference materials were prepared in a glove-box module consisting of a cube-shaped glass box isolated from the ambient temperature and 40 μl of HU solution in deionized water at different concentrations were added to reproduce the potential amounts of HU on a filter (50, 3.5 ng, and 50 μg).

## MicroNIR/Chemometrics Method

Spectra were collected by a portable, ultra-compact and low-cost device MicroNIR spectrometer, developed and distributed by Viavi Solutions (JDSU Corporation, Milpitas, USA). This device operates in the spectral region 900–1,700 nm and consists of a linear variable filter (LVF) as dispersing element directly connected to a 128-pixel linear indium gallium arsenide (InGaAs) array detector and two tungsten light bulbs as radiation source.

In the MicroNIR, measuring the optimum focal point of the illumination source from the spectrometer's window to a sample is achieved by the means of a special collar. As a result, this particular geometry permits to achieve comparable outcomes as the reduction in size does not compromise the performances of the spectrometer. The instrument control was performed by the MicroNIR Pro software (JDSU Corporation, Milpitas, USA) and chemometric tools such as Principal Component Analysis (PCA) and Partial Least Square (PLS) algorithms were used as unsupervised technique and calibration models by V-JDSU Unscrambler Lite (Camo software AS, Oslo, Norway).

Spectra were collected at a nominal spectral resolution of 6.25 nm in the reflectance mode. Spectralon was used as NIR reflectance standard (blank), with a 99% diffuse reflectance, while a dark reference was obtained from a fixed place in the room. The acquisitions were performed with an integration time of 10 ms, resulting in a total measurement time of 2.5 s for each sample.

As recommended for spectroscopic data (Rinnan et al., 2009), mathematical pre-treatments were considered for chemometric evaluation such as scatter-correction methods [Standard Normal Variate transform (SNV) (Barnes et al., 1989), Multiplicative Scatter Correction (MSC) (Geladi et al., 1985), and Mean Centering (Wold and Sjöström, 1977)], Savitzky-Golay (SG) polynomial derivative filters (Savitzky-Golay, 1964) as spectral derivation techniques. Among these pre-treatments, the combination of second derivative algorithm followed by Mean Centering was selected because it provided the best outcomes in terms of Root Mean-Squared Error of Calibration (RMSEC), Root Mean-Squared Error of Prediction (RMSEP), and coefficient of determination ($R^2$) (Miller and Miller, 2000; Mark and Workman, 2007).

Figures of merits were used to estimate model performances. In particular, Residual Predictive Deviation (RPD) was used to evaluate correction forecasting model and calculated as the standard deviation (*SD*)/RMSEP. In general, the model is considered stable when RPD $\geq$3 or not satisfactory when RPD $<$2. In this work, the precision of the method was determined on nine different samples with concentrations regularly

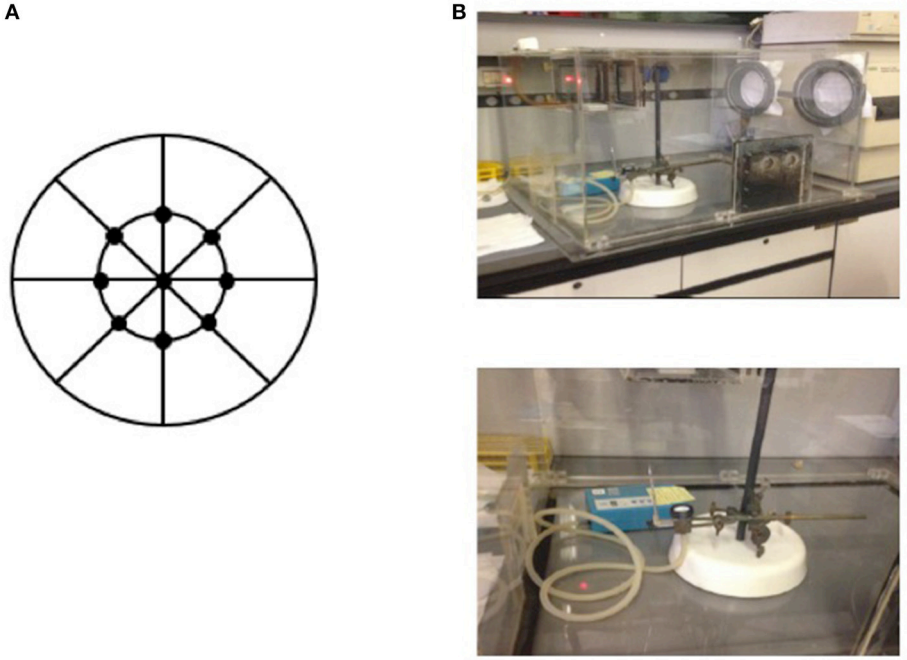



**FIGURE 1 |** Sampling procedure of HU on a filter by acquiring nine spectra for each membrane **(A)** in a preserved glove box **(B)**.

distributed along the linear range, using nine replicates in the same day.

Sensitivity (SEN) represents the fraction of the analytical signal responsible for an increase in the concentration of HU and was calculated as follows: SEN = 1/b, where b is the vector of regression coefficients with A latent variables. The minimum detectable concentration (MDC) is defined as the lowest concentration that can be reliably measured according to ISO 11843-2:2000 recommendations[1].

## Experimental Design

Calibration and validation models were developed using the dataset from 297 samples. The data set was divided into two groups, the calibration set (216 samples) and validation set (81 samples). In order to provide a sample selection for the calibration and validation set as representative as possible and to ensure uniformity of dataset, the X and Y distances were taken into account simultaneously, by applying the Kennard–Stone (KS) uniform sampling algorithm. The calibration set consisted of a series of reference samples including blanks (filters without HU) and fortified blanks with increasing amounts of HU (50, 3.5 ng, and 50 μg).

A comprehensive sampling procedure was scheduled as follows: samples were collected in a preserved glove box and nine spectra were acquired in reflectance mode for each membrane, as shown in **Figure 1**. A total of nine filters were used to optimize the model of prediction for HU exposure. Six independent batches were prepared for calibration; while validation was performed on the same type of samples as the calibration set, but fully independent batches, using three series of filters.

## GC-MS Method

GC-MS analysis was done on a Perkin Elmer system (Waltham, MA) using a HP-5MS (30 m × 0.25 mm × 0.25 mm) as capillary separation column. Electron impact (EI) ionization was employed at a voltage of 70 eV. The carrier gas was helium delivered at a constant flow of 1 mL/min. The oven temperature program was initially set at 150°C for 1 min, ramped to 140°C at 12°C/min and maintained for 1 min, and then ramped to 270°C at 35°C/min for 2.5 min. The temperatures for the inlet, interface, ion source and quadrupole were set at 270, 250, 230, and 150°C, respectively. Mass spectral data was collected in the scan mode from m/z 44 to 400; in the SIM mode, fragments at 277 and 292 m/z were monitored for quantification and confirmation, respectively.

# RESULTS

To develop a novel analytical method to monitor occupational exposure to cancerogenic agents by evaluating the amount of HU on a filter, multivariate statistical analysis was performed for optimal selection of the experimental procedure. As a consequence, a number of variables were considered in order to ensure a correct and representative sampling procedure: (a) membrane type and sampling side; (b) sampling procedure to reproduce HU exposure in terms of volume to be added on a filter; (c) spectra acquisition. Preliminarily,

[1]ISO 11843-2:2000. Capability of Detection, International Standards Organization. Geneva.

all the acquired MicroNIR data corresponding to different experimental conditions were pre-treated and processed by a simple exploratory tool such as PCA. After that, a prediction model of HU based on Partial Least Square Regression (PLSR) was entirely developed and validated.

## Sampling Procedure Optimization

To make the method representative, the first investigated issue consisted of reference material preparation. Two different ways of HU deposition on a filter were investigated: (i) calibration on different filters i.e., four different filters (one blank and three fortified blanks) were considered; and (ii) calibration on a single filter i.e., only one filter was used and progressively fortified with increasing amounts of HU. In this case, spectra of blank and fortified samples were acquired prior to each deposition by the portable MicroNIR. In the first case, samples were prepared using 40 µl of aqueous solution of HU on each filter; while in the second case, a volume of 15 µl was used for each deposition.

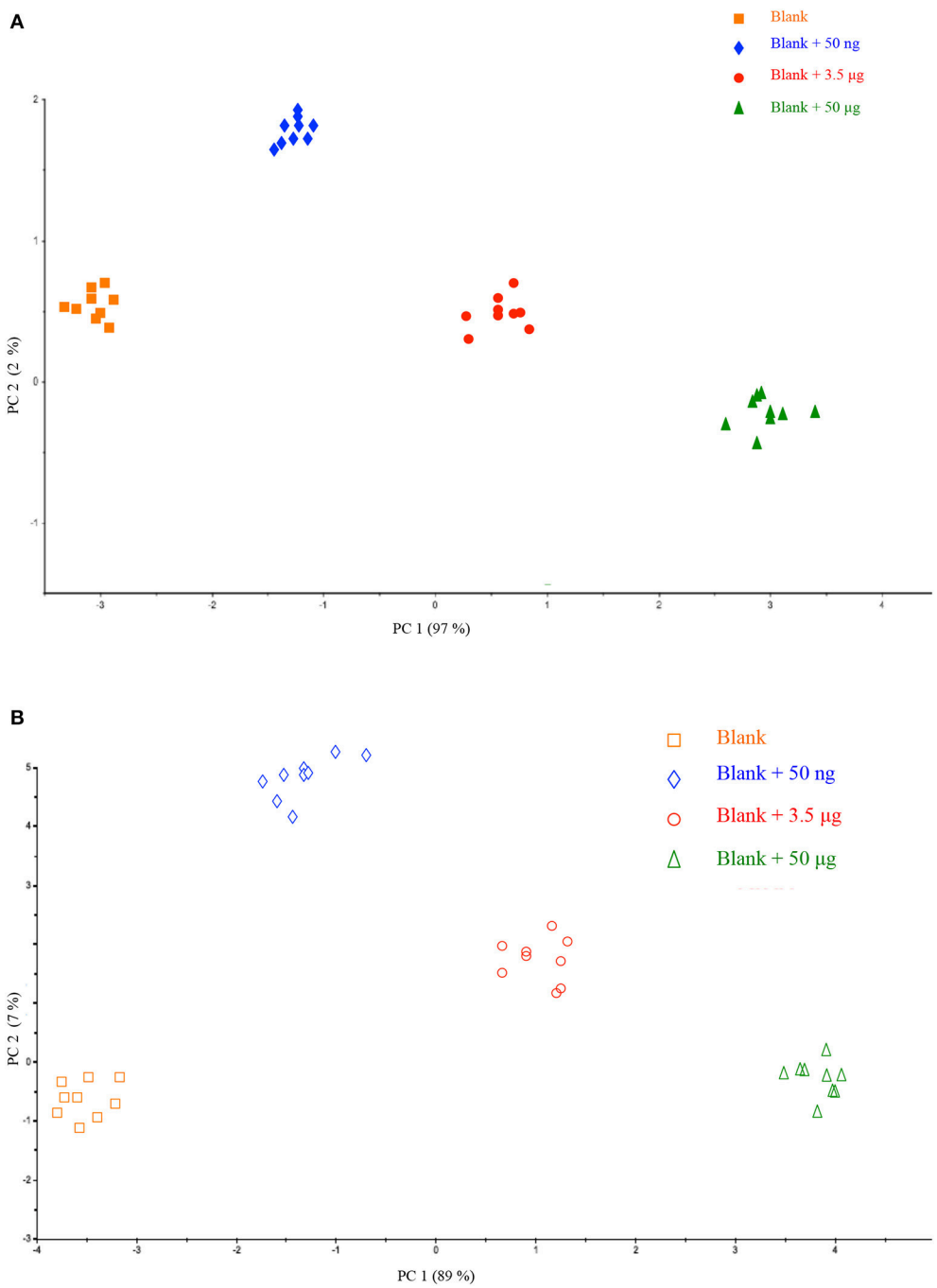



**FIGURE 2 |** Scores plot of the Principal Component Analysis performed on the dataset related to calibration on different filters **(A)** and calibration on a single filter **(B)**.

All the acquired spectra were pre-treated and analyzed simultaneously by PCA. As displayed in **Figure 2**, each point represents an average of the nine respective spectra of a filter and colors were used to highlight the quantity of HU. The interpretation of the scores plot provides preliminary important information with respect to HU deposition and correlation to its different amounts on a filter.

A good correlation could be observed for samples of the same class (blank and fortified blanks) as there was no data dispersion, suggesting a correct repeatability of the method. This observation is very interesting because it is possible to clearly discriminate HU quantity on the membrane of a filter. For any deposition way, hence, the method would be suitable in practice where occupational exposure of workers may be monitored by a personal sampling system collecting a real blank (prior to HU handling) to be fortified and directly analyzed.

In addition, as shown in **Figure 2**, in both cases moving along PC1 (97 and 89% of explained variance) all the analyzed samples could be well grouped according to HU amount. It further confirms the ability of the approach MicroNIR/Chemometrics in monitoring occupational exposure to HU according to its amount collected on a filter.

A deeper investigation of the acquired spectra was performed by comparing the two series of samples (four- and one-filter calibration) in a single dataset. **Figure 3** displays PCA data showing that the same samples can be divided into two main groups according to PC2: four- and one-filter calibration. As a result of the PCA data, the different locations of samples in the plot indicate the contribution of HU deposition way on the spectroscopic signal.

Such a result is not surprising when a reflectance acquisition mode is involved, because the surface of the filter membrane may have some influence on the spectral response as a function of the volume added. Despite the different behaviors, samples could be clearly differentiated according to PC1 (91% of explained variance) and the preliminary outcomes suggested the possibility to further investigate the repeatability of the method.

With the aim of extending this procedure to real samples, nine different filters were prepared and subsequently fortified with different amounts of HU so as to increase the number of investigated samples and evaluate whether the method would be batch-dependent. As shown in **Figure 4**, all the samples of the same class (displayed in different colors), could be well grouped and located in the plot according to PC1. In addition, no dispersion of data was observed thus indicating the effectiveness of the optimized HU deposition on a filter. On the basis of preliminary interesting results, a prediction model of HU on a single filter membrane was successfully validated.

## PLS Model of Prediction

In order to obtain the best results of calibration, the effect of a number of pre-treatments was evaluated i.e., the combination of spectral pre-treatments and wavelength range selection.

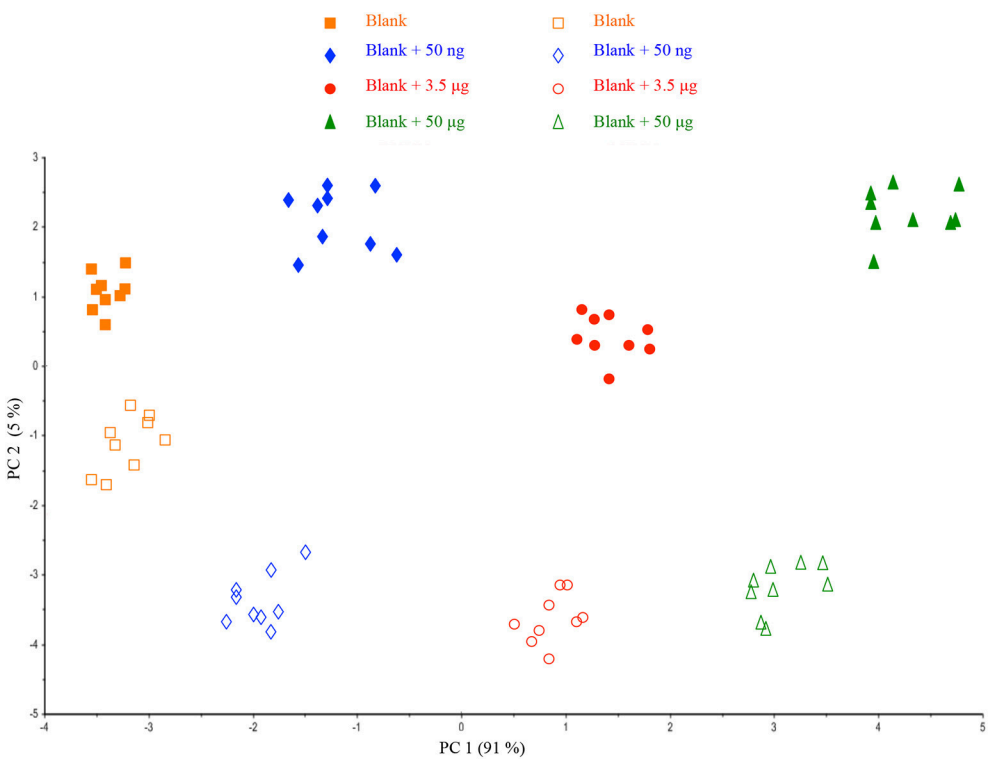



**FIGURE 3 |** Scores plot of the Principal Component Analysis performed on the datasets from the two different ways of HU deposition on a filter.

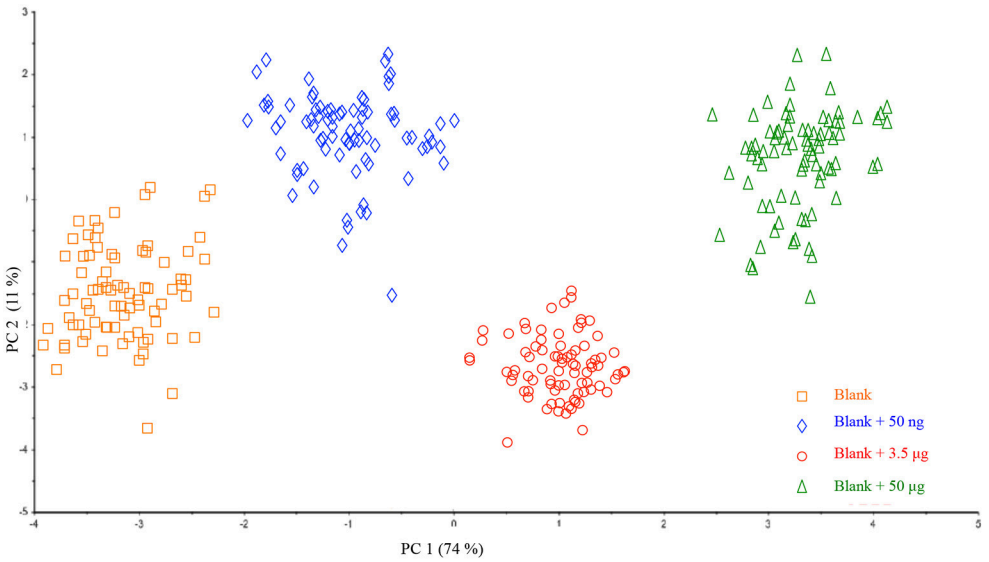



**FIGURE 4 |** Scores plot of the Principal Component Analysis performed on the entire dataset of collected samples.

**TABLE 1 |** Figures of merit of HU calculated with different spectral pre-treatments in calibration and prediction steps.

| Pre-treatment | Calibration | | Prediction | | |
|---|---|---|---|---|---|
| | $R^2$ | RMSEC | $R^2$ | RMSEP | RPD |
| SNV + Mean centering | 0.9985 | 1.98 | 0.9973 | 2.14 | 1.9 |
| MSC + Mean centering | 0.9889 | 2.02 | 0.9817 | 1.26 | 1.5 |
| 1st derivative + Mean centering | 0.9999 | 0.61 | 0.9998 | 1.02 | 2.1 |
| 2nd derivative + Mean centering | 1.0000 | 0.09 | 1.0000 | 0.12 | 5.4 |

Calibration and validation sets were pre-processed using Standard Normal Variate (SNV) scaling (Barnes et al., 1989), MSC (Geladi et al., 1985), and Mean Centering (Wold and Sjöström, 1977), Savitzky-Golay (SG) polynomial derivative filters (Savitzky-Golay, 1964) and a combination of these pre-treatments.

For evaluation of model performances, comparison was made for different spectral pre-treatments to identify the most effective one in terms of prediction error using the Predicted Residual Error Sum of Squares (PRESS) to represent the sum of squares of the prediction error and the coefficient of determination ($R^2$). Usually, the smaller the PRESS value is, the better the model's predictive ability is. $R^2$ provides the percentage variation in y explained by x-variables and is largely used to evaluate the fitting performance. Satisfactory results ($R^2$ and RMSEC) were obtained for the calibration of HU as shown in **Table 1**.

Good model agreement is confirmed in the validation step ($R^2 > 0.9817$ and RMSEP $< 2.14$ for all the optimized models). As far as the data are concerned, the best performance can be achieved by using second derivative pre-treatment followed by mean centering (4 latent variables) as it provides the lowest RMSE and highest $R^2$ values. Furthermore, the effect of the variable spectral selection within the calibration block was evaluated to improve the model's ability to predict HU. As illustrated in **Figure 5**, the first principal component loadings accounted for more than 87% of the total variance.

Validation results of the most performing model (second derivative pre-treatment followed by mean centering) after variable selection in the range 1,540–1,600 nm are reported in **Table 2**, showing that the optimized model could quantify HU on a glass fiber filter with limit of detection of 50 ng/filter. This finding points out that an adequate PLS regression model can help quantify HU directly from MicroNIR measurements without any prior sample preparation.

## Evaluation of Prediction Ability

The validated model was consequently used to process 30 filters collected during routine HU handling. In order to evaluate the prediction ability of the model, all the samples were simultaneously analyzed by the reference method (GC-MS) and MicroNIR/Chemometrics approach. Data obtained from the MicroNIR approach (**Table 3**) show that the PLS model permitted to achieve the best prediction precision with RMSEP of 0.12 and RPD of 6.1, which ensured the accuracy and robustness of the model.

In addition, the chromatographic analysis detected HU in only 19 of the 30 samples as the Limit of Detection (LOD) and Limit of Quantification (LOQ) of this method were 0.7 and 2.5 µg, respectively. When the amount of HU was chromatographically found to be lower than the LOQ of the method, LOD was used to compare with the predicted values obtained by MicroNIR/Chemometrics approach. The results showed a $R^2$ of 0.99 and acceptable values of bias at 95% confidence (see **Table 3**).

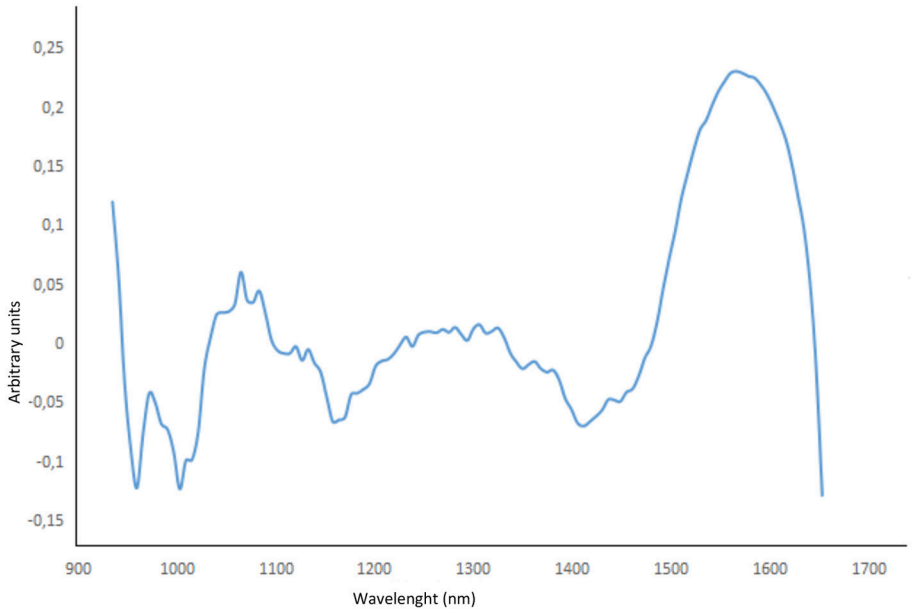



**FIGURE 5 |** First principal component loading variation and variable spectral selection.

**TABLE 2 |** Analytical figures of merit for PLS quantification model.

| **Figures of merit** | |
|---|---|
| RMSEC | 0.09 |
| RMSEP | 0.12 |
| RPD | 5.4 |
| LV* | 4 |
| $R^2$ Validation | 1.000 |
| Precision | 1.24 |
| Sensitivity $(\%w/w)^{-1}$ | 0.100 |
| MDC** (ng) | 50 |
| Range (μg) | 0.05-50 |
| Mean ± *SD* (μg) | 23.8 ± 0.65 |

**Latent variables. **Minimum detection concentration.*

**TABLE 3 |** Results of the MicroNIR approach.

| **Figures of merit** | |
|---|---|
| RMSEP | 0.12 |
| RPD | 6.1 |
| Slope | 0.990 |
| Bias | 0.016 |
| Range (μg) | 0.08–42.8 |
| Mean ± *SD* (μg) | 3.6 ± 0.73 |

MicroNIR computed values were found to be significantly lower than corresponding GC ones as the LOD of the MicroNIR method is 50 ng, meaning that the MicroNIR/Chemometrics can be a promising approach for occupational exposure monitoring at HU low levels.

## CONCLUSIONS

An ultra-compact portable device (MicroNIR) was applied to assess a novel way for HU occupational exposure monitoring.

A comprehensive sampling procedure was pointed out. Chemometric evaluation of spectra collected by a miniaturized device operated in the Near Infrared region, was optimized and entirely validated by PLS regression. The proposed method has the advantage of simplicity and avoiding sample pre-treatment, thus limiting even the analyst's HU exposure. Moreover, this approach may be considered as the optimal technology to determine cancerogenic agents or other dangerous molecules in a single-touch analysis as it is entirely portable and non-destructive. The achieved results highlight the extremely high potential of MicroNIRs to detect the HU with lower detection limits with respect to reference methods. To the best of the authors' knowledge, this approach would be the first ever proposed for the on-site detection of HU. It requires no sample preparation, is non-destructive and easy to perform (no highly-skilled personnel required), allowing a rapid evaluation of the HU occupational exposure.

## AUTHOR CONTRIBUTIONS

SM and RR conceived the study and developed the experimental design. RR performed the chemometric evaluation of data. SM and RR analyzed and interpreted data and wrote the manuscript. All authors reviewed and approved the manuscript.

## REFERENCES


Barnes, R. J., Dhanoa, M. S., and Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectros*. 43, 772–777. doi: 10.1366/00037028942 02201

Bolton, B. H., Woods, L. A., Kaung, D. T., and Lawton, R. L. (1965). A simple method of colorimetric analysis for hydroxyurea (NSC-32065). *Cancer Chemother. Rep*. 46, 1–5.

Broto, M., Galve, R., and Marco, M. P. (2017). Bioanalytical methods for cytostatic therapeutic drug monitoring and occupational exposure assessment. *Trends Analyt. Chem*. 93, 152–170. doi: 10.1016/j.trac.2017.05.005

Candrilli, S. D., O'Brien, S. H., Ware, R. E., Nahata, M. C., Seiber, E. E., and Balkrishnan, R. (2011). Hydroxyurea adherence and associated outcomes among Medicaid enrollees with sickle cell disease. *Am. J. Hematol*. 86, 273–277. doi: 10.1002/ajh.21968

Dalton, R. N., Turner, C., Dick, M., Height, S. E., Awogbade, M., Inusa, B., et al. (2005). The measurement of urinary hydroxyurea in sickle cell anaemia. *Br. J. Haematol*. 130, 138–144. doi: 10.1111/j.1365-2141.2005.05583.x

Davidson, J. D., and Winter, T. S. (1963). A method of analyzing for hydroxyurea in biological fluids. *Cancer Chemother. Rep*. 27, 97–110.

Davies, S. C., and Gilmore, A. (2003). The role of hydroxyurea in the management of sickle cell disease. *Blood Rev*. 17, 99–109. doi: 10.1016/S0268-960X(02)00074-7

De Marco, R., Di Gioia, M. L., Liguori, A., Perri, F., , Siciliano, C., and Spinella, M. (2011). N alkylation of N-arylsulfonyl-α-amino acid methyl esters by trialkyloxonium tetrafluoroborates. *Tetrahedron* 67, 9708–9714. doi: 10.1016/j.tet.2011.10.042

Elchuri, S. V., Williamson, R. S., Brown, R. C., Haight, A. E., Spencer, J. B., Buchanan, I., et al. (2015). The effects of hydroxyurea and bone marrow transplant on Anti-Müllerian hormone (AMH) levels in females with sickle cell anemia. *Blood Cells Mol. Dis*. 55, 56–61 doi: 10.1016/j.bcmd.2015.03.012

Flanagana, J. M., Howarda, T. A., Mortier, N., Avlasevichb, S. L., Smeltzer, M. P., Wu, S., et al. (2010). Assessment of genotoxicity associated with hydroxyurea therapy in children with sickle cell anemia. *Mutat. Res. Genet. Toxicol. Environ. Mutagen*. 698, 38–42 doi: 10.1016/j.mrgentox.2010.03.001

Garg, U., Scott, D., Frazee, C., Kearns, G., and Neville, K. (2015). Isotope-dilution gas chromatography-mass spectrometry method for the analysis of hydroxyurea. *Ther. Drug Monit*. 37, 325–330. doi: 10.1097/FTD.0000000000000145

Geladi, P., MacDougall, D., and Martens, H. (1985). Linearization and scatter-correction for Near-Infrared reflectance spectra of meat. *Appl. Spectrosc*. 39, 491–500. doi: 10.1366/0003702854248656

Hai, X., Guoa, M., Gaoa, C., and Zhoub, J. (2017). Quantification of hydroxyurea in human plasma by HPLC–MS/MS and its application to pharmacokinetics in patients with chronic myeloid leukaemia. *J. Pharm. Biomed. Anal*. 137, 213–219. doi: 10.1016/j.jpba.2017.01.008

Heeney, M. M., and Ware, R. E. (2008). Hydroxyurea for children with sickle cell disease. *Pediatr. Clin. N. Am*. 55, 483–501. doi: 10.1016/j.pcl.2008.02.003

Italia, K., Jain, D., Gattani, S., Jijina, F., Nadkarni, A., Sawant, P., et al. (2009). Hydroxyurea in sickle cell disease-A study of clinico-pharmacological efficacy in the Indian haplotype. *Blood Cells Mol. Dis*. 42, 25–31. doi: 10.1016/j.bcmd.2008.08.003

Iyamu, E. W., Roa, P. D., Kopsombut, P., Aguinaga, M. D., and Turner, E. A. (1998). New isocratic high-performance liquid chromatographic procedure to assay the anti-sickling compound hydroxyurea in plasma with ultraviolet detection. *J. Chromatogr. B Biomed. Sci. Appl*. 709, 119–126. doi: 10.1016/S0378-4347(98)00020-6

James, H., Nahavandi, M., Wyche, M. Q., and Taylor, R. E. (2006). Quantitative analysis of trimethylsilyl derivative of hydroxyurea in plasma by gas chromatography-mass spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci*. 831, 42–47. doi: 10.1016/j.jchromb.2005.11.033

Karsy, M., Hoang, N., Barth, T., Burt, L., Dunson, W., Gillespie, D. L., et al. (2016). Combined hydroxyurea and verapamil in the clinical treatment of refractory meningioma: human and orthotopic xenograft studies. *World Neurosurg*. 86, 210–219. doi: 10.1016/j.wneu.2015.09.060

Kettani, T., Cotton, F., Gulbis, B., Ferster, A., and Kumps, A. (2009). Plasma hydroxyurea determined by gas chromatography-mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci*. 87, 446–450. doi: 10.1016/j.jchromb.2008.12.048

Legranda, T., Rakotoson, M. G., Galactérosb, F., Bartoluccib, P., and Hulina, A. (2017). Determination of hydroxyurea in human plasma by HPLC-UV using derivatization with xanthydrol. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci*. 1064, 85–91. doi: 10.1016/j.jchromb.2017.09.008

Liew, L. P., Lim, Z. Y., Cohen, M., Kong, Z., Marjavaara, L., Chabes, A., et al. (2016). Hydroxyurea-mediated cytotoxicity without inhibition of ribonucleotide reductase. *Cell Rep*. 17, 1657–1670. doi: 10.1016/j.celrep.2016.10.024

Lori, F., Malykh, A., Cara, A., Sun, D., Weinstein, J. N., Lisziewicz, J., et al. (1994). Hydroxyurea as an inhibitor of human immunodeficiency virus-type 1 replication. *Science* 266, 801–805. doi: 10.1126/science.7973634

Main, K. B., Medwick, T., Bailey, L. C., and Shinkai, J. H. (1987). Quantitative analysis of hydroxyurea and urea by proton nuclear magnetic resonance (NMR) spectroscopy. *Pharm. Res*. 4, 412–415. doi: 10.1023/A:10164904 30135

Manouilov, K. K., McGuire, T. R., and Gwilt, P. R. (1998). Colorimetric determination of hydroxyurea in human serum using high-performance liquid chromatography. *J. Chromatogr. B Biomed. Sci. Appl*. 708, 321–324. doi: 10.1016/S0378-4347(97)00634-8

Marahatta, A., Megaraj, V., McGann, P. T., Ware, R. E., and Setchell, K. D. (2016). Stable-isotope dilution HPLC-electrospray ionization tandem mass spectrometry method for quantifying hydroxyurea in dried blood samples. *Clin. Chem*. 62, 1593–1601. doi: 10.1373/clinchem.2016.263715

Marahatta, A., and Ware, R. E. (2017). Hydroxyurea: analytical techniques and quantitative analysis. *Blood Cells Mol. Dis*. 67, 132–142. doi: 10.1016/j.bcmd.2017.08.009

Mark, H., and Workman, J. (2007). *Chemometrics in Spectroscopy*. Amsterdam: Elsevier/Academic Press.

Materazzi, S., Gregori, A., Ripani, L., Apriceno, A., and Risoluti, R. (2017a). Cocaine profiling: implementation of a predictive model by ATR-FTIR coupled with chemometrics in forensic chemistry. *Talanta* 166, 328–335. doi: 10.1016/j.talanta.2017.01.045

Materazzi, S., Peluso, G., Ripani, L., and Risoluti, R. (2017b). High-throughput prediction of AKB48 in emerging illicit products by NIR spectroscopy and chemometrics. *Microchem. J*. 134, 277–283. doi: 10.1016/j.microc.2017. 06.014

Materazzi, S., Risoluti, R., Pinci, S., and Saverio Romolo, F. (2017c). New insights in forensic chemistry: NIR/Chemometrics analysis of toners for questioned documents examination. *Talanta* 174, 673–678. doi: 10.1016/j.talanta.2017.06.044

Milks, J. E., and Janes, R. H. (1956). Separation and detection of cyanamide and its derivatives and determination of urea by paper chromatography. *Anal. Chem*. 28, 846–849. doi: 10.1021/ac60113a019

Miller, J. N., and Miller, J. C. (2000). *Statistics and Chemometrics for Analytical Chemistry*. Harlow: Prentice Hall.

Millicovsky, G., DeSesso, J. M., Kleinman, L. I., and Clarka, K. E. (1981). Effects of hydroxyurea on hemodynamics of pregnant rabbits: a maternally mediated mechanism of embryotoxicity. *Am. J. Obstet. Gynecol*. 140, 747–752. doi: 10.1016/0002-9378(81)90734-1

Naik, K. M., Ashi, C. R., and Nandibewoor, S. T. (2015). Anodic voltammetric behavior of hydroxyurea and its electroanalytical determination in pharmaceutical dosage form and urine. *J. Electroanal. Chem*. 755, 109–114. doi: 10.1016/j.jelechem.2015.07.038

Naumann, B. D., Sargent, E. V., Starkman, B. S., Fraser, W. J., Becker, G. T., and Kirk, G. D. (1996). Performance-based exposure control limits for pharmaceutical active ingredients. *Am. Ind. Hyg. Assoc*. 57, 33–42. doi: 10.1080/15428119691015197

Oliveri, P., Di Egidio, V., Woodcock, T., and Downey, G. (2011). Application of class modelling techniques to near infrared data for food authentication purposes. *Food Chem*. 125, 1450–1456. doi: 10.1016/j.foodchem.2010. 10.047

Osytek, A., Biesaga, M., Pyrzynska, K., and Szewczynska, M. (2008). Quantification of some active compounds in air samples at pharmaceutical workplaces by HPLC. *J. Biochem. Biophys. Methods* 70, 1283–1286. doi: 10.1016/j.jbbm.2007.10.003

Pujari, M. P., Barrientos, A., Muggia, F. M., and Koda, R. T. (1997). Determination of hydroxyurea in plasma and peritoneal fluid by high-performance liquid

chromatography using electrochemical detection. *J. Chromatogr. B Biomed. Sci. Appl*. 694, 185–191. doi: 10.1016/S0378-4347(97)00120-5

Quattrone, F., Dini, V., Barbanera, S., Zerbinati, N., and Romanelli, M. (2013). Cutaneous ulcers associated with hydroxyurea therapy. *J Tissue Viab.* 22, 112–121. doi: 10.1016/j.jtv.2013.08.002

Rinnan, A., Van den Berg, F., and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends Analy. Chem*. 28, 1201–1222. doi: 10.1016/j.trac.2009.07.007

Risoluti, R., Gregori, A., Schiavone, S., and Materazzi, S. (2018). "click and screen" technology for the detection of explosives on human hands by a portable MicroNIR-chemometrics platform. *Anal. Chem*. 90, 4288–4292. doi: 10.1021/acs.analchem.7b03661

Risoluti, R., Materazzi, S., Gregori, A., and Ripani, L. (2016a). Early detection of emerging street drugs by near infrared spectroscopy and chemometrics. *Talanta* 153, 407–413. doi: 10.1016/j.talanta.2016.02.044

Risoluti, R., Materazzi, S., Sorrentino, F., and Caprari, P. (2016b). Thermogravimetric analysis coupled with chemometrics as a powerful predictive tool for ß-thalassemia screening. *Talanta* 159, 425–432. doi: 10.1016/j.talanta.2016.06.037

Savitzky-Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem*. 36, 1627–1639. doi: 10.1021/ac60214a047

Sivakumar, P., Meenakshi, S., Govindan, P., and Subba Rao, R. V. (2013). Spectrophotometric determination of hydroxyurea and stability in nitric acid medium. *Int. J. Nucl. Energy Sci. Eng*. 3, 27–31.

Sorg, B. L., Hull, W. E., Kliem, H. C., Mierc, W., and Wiessler, M. (2005). Synthesis and NMR characterization of hydroxyurea and mesylglycol glycoconjugates as drug candidates for targeted cancer chemotherapy. *Carbohydr. Res.* 340, 181–189. doi: 10.1016/j.carres.2004.11.024

Spivak, J. L., and Hasselbalch, H. (2011). Hydroxycarbamide: a user's guide for chronic myeloproliferative disorders. *Expert Rev. Anticancer. Ther*. 11, 403–414. doi: 10.1586/era.11.10

Usawanuwat, J., Boontanon, N., and Boontanon, S. K. (2014). Analysis of three anticancer drugs (5-fluorouracil, cyclophosphamide and hydroxyurea) in water samples by HPLC MS/MS. *Int. J. Adv. Agr. Environ. Eng*. 1, 72–76. doi: 10.15242/IJCCIE.C0114136

Wold, S., and Sjöström, M. (1977). "SIMCA: A method for analyzing chemical data in terms of similarity and analogy," in *Chemometrics: Theory and Application, Vol. 12*, ed B. R. Kowalski (Washington, DC: CS Symposium Series), 243–242. doi: 10.1021/bk-1977-0052.ch012

Woo, G. H., Bakb, E. J., Nakayama, H., and Doi, K. (2005). Hydroxyurea (HU)-induced apoptosis in the mouse fetal lung. *Exp. Mol. Pathol*. 79, 59–67. doi: 10.1016/j.yexmp.2005.02.007

Yarbro, J. W., and Leavell, U. W. (1969). Hydroxyurea: a new agent for the management of refractory psoriasis. *J. Ky. Med. Assoc*. 67, 899–901.

Zontova, Y. V., Balyklova, K. S., Titovac, A. V., Rodionova, O. Y., and Pomerantsev, A. L. (2016). Chemometric aided NIR portable instrument for rapid assessment of medicine quality. *J. Pharm. Biomed. Anal*. 131, 87–93. doi: 10.1016/j.jpba.2016.08.008

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 28 June 2018
doi: 10.3389/fchem.2018.00261

# A Plasma Biochemical Analysis of Acute Lead Poisoning in a Rat Model by Chemometrics-Based Fourier Transform Infrared Spectroscopy: An Exploratory Study

*Wenli Tian, Dan Wang, Haoran Fan, Lujuan Yang and Gang Ma**

*Key Laboratory of Medicinal Chemistry and Molecular Diagnosis of Ministry of Education, Key Laboratory of Analytical Science and Technology of Hebei Province, College of Chemistry and Environmental Science, Hebei University, Baoding, China*

In this work, we explored to use chemometrics-based Fourier transform infrared (FTIR) spectroscopy to investigate the plasma biochemical changes due to acute lead poisoning (ALP) in a rat model. We first collected the FTIR spectra of the plasma samples from the rats with and without suffering from ALP. We then performed the chemometric analysis of these FTIR spectra using principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). We found that the chemometrics-based FTIR spectroscopy can discriminate the rats with and without ALP. Further analysis on the PLS-DA regression coefficient revealed that the spectral changes, in particular, corresponding to the biochemical changes of proteins in the plasma may be used as potential spectral biomarkers for the diagnostics of lead poisoning. Our work demonstrates the potential of chemometrics-based FTIR spectroscopy as a promising tool for the biochemical analysis of plasma that could consequently enable an objective, convenient and non-destructive diagnostics of lead poisoning. To the best of our knowledge, this work is the first application of chemometrics-based FTIR spectroscopy in the diagnostics of lead poisoning.

**Keywords: FTIR spectroscopy, infrared spectroscopy, chemometrics, lead poisoning, acute lead poisoning, principle component analysis, partial least squares discriminant analysis**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Zoltán Kónya,*
*University of Szeged, Hungary*
*Chih-Ching Huang,*
*National Taiwan Ocean University, Taiwan*

****Correspondence:***
*Gang Ma*
*gangma@hbu.edu.cn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *29 March 2018*
***Accepted:*** *11 June 2018*
***Published:*** *28 June 2018*

***Citation:***
*Tian W, Wang D, Fan H, Yang L and Ma G (2018) A Plasma Biochemical Analysis of Acute Lead Poisoning in a Rat Model by Chemometrics-Based Fourier Transform Infrared Spectroscopy: An Exploratory Study. Front. Chem. 6:261. doi: 10.3389/fchem.2018.00261*

## INTRODUCTION

Lead is an omnipresent metal that has been used since prehistoric times. Prior to the industrial revolution, human exposure to lead in the environment was relatively low, but significantly increased over time due to modern industrial activities. It is estimated that over 300 million tons of lead has been released to the environment by human activities (Tong et al., 2000), which leads to a rapid increase in lead exposure to the environment. A previous study indicated that the lowest levels of human blood lead in industrial era were 50–200 times higher than preindustrial era (Flegal and Smith, 1992b). As for lead poisoning, in 1839, Tanqueral des Planches described the symptoms of acute lead poisoning (ALP) and studied the signs of ALP in adults (Hunter, 1978). In the middle and late nineteenth century, lead poisoning became a serious health problem among

Britain workers. British Parliament eventually enacted relevant laws and regulations to prevent lead poisoning (Hunter, 1978; Smith, 1984; Winder, 1984; Tong et al., 2000). Lead poisoning can be caused by human ingestion and respiration of lead and related products such as lead-containing paints. Lead can cause a series of physiological and biochemical changes within human body, affecting central and peripheral nervous system, cardiovascular system, reproductive system, immune system, gastrointestinal tract, liver, kidney and brain (Hunter, 1978; Smith, 1984; Winder, 1984; Kazantzis, 1989; Goldstein, 1992; Tong, 1998; Tong et al., 2000).

The basic principle in lead poisoning diagnostics is based on the determination of lead level in human body. There are currently several methods available for measuring lead in blood samples. For example, one common method is the so-called blood film method, in which the morphology of the red blood is examined with a microscope to reveal basophilic stippling of red blood cells (i.e., red blood cells with dots in their morphologies). However, this method is not very specific because other unrelated conditions (such as folate and vitamin B12 deficiencies) can also give basophilic stippling of red blood cells. Lead level can be evaluated indirectly by measuring erythrocyte protoporphyrin (EP) in blood samples. It is noted that such EP measurement is not very sensitive and specific because an increase in EP level can also be observed in the case of iron deficiency. X-ray fluorescence method can be used to determine the cumulative exposure and total body burden of lead. However, this method is not so convenient because X-ray fluorescence instrument is not widely available in clinic. Apparently, the current methods in lead poisoning diagnostics still have some limitations and disadvantages (Patrick, 2006; Brodkin et al., 2007). Searching for a specific, rapid, convenient, objective and cost-effective method for lead poisoning diagnostics is no doubt very meaningful (Flegal and Smith, 1992a).

In recent years, Fourier transform infrared (FTIR) spectroscopy has been widely used in the biochemical analysis field (Baker et al., 2014). FTIR spectroscopy is a simple, convenient, non-destructive, rapid and low-cost detection method to sample biological materials such as blood and tissue for diagnostic purposes (Deleris and Petibois, 2003; Ellis and Goodacre, 2006; Krafft et al., 2007, 2009; Gasper et al., 2009; Gajjar et al., 2013; Baker et al., 2014; Mitchell et al., 2014; Ollesch et al., 2014; Sheng et al., 2015; Staniszewska-Slezak et al., 2015; Depciuch et al., 2017; Elmi et al., 2017; Ghimire et al., 2017; Guo et al., 2017; Le Corvec et al., 2017; Li et al., 2017; Liu et al., 2017; Paraskevaidi et al., 2017; Roy et al., 2017; Sarkar et al., 2017; Titus et al., 2017; De Bruyne et al., 2018; Rai et al., 2018). When combined with chemometric analysis, FTIR spectroscopy can be further empowered in disease diagnostics. Now, FTIR spectroscopy has been used in many studies to detect the physiological states and disease-specific biomarkers in the blood. For example, Staniszewska-Slezak et al. established the rat models for pulmonary arterial hypertension and systemic hypertension, and then collected the FTIR spectra of rat plasma samples. By using FTIR spectroscopy combined with principal component analysis (PCA), they found that they could distinguish the two different hypertension states as well as the healthy state. They also envisioned that chemometrics-based FTIR spectroscopy could potentially provide some spectral biomarkers for disease diagnostics (Staniszewska-Slezak et al., 2015). Roy et al. recently used attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy in combination with partial least squares discriminant analysis and partial least squares regression to identify malaria parasites, blood glucose and urea levels in whole blood samples (Roy et al., 2017). Titus et al. recently proposed an FTIR approach combined with cluster and heterogeneity analyses to rapidly screen colitis without using biopsies or *in vivo* measurements (Titus et al., 2017). Paraskevaidi et al. recently demonstrated an excellent diagnostic performance of chemometrics-based ATR-FTIR spectroscopy by analyzing plasma samples from patients with Alzheimer's disease (Paraskevaidi et al., 2017).

In our work, we focused on the biochemical changes of plasma after lead poisoning using a rat model suffering from ALP. The main goal of this study was to find the plasma biochemical changes induced by lead in rats by FTIR spectroscopy combined with chemometric approaches such as PCA and partial least squares discriminant analysis.

# EXPERIMENTAL

## ALP Rat Model

Male Wister rats (240 $\pm$ 20 g) were purchased from the Vital River Lab Animal Technology Co., Ltd. (Beijing, China). Animals were housed under constant temperature, humidity and lighting (12 h per day) and were allowed free access to food and water. The animal experiment was carried out in accordance with the guidelines for the care and use of laboratory animals and the relevant ethical regulations of the Animal Ethics Committee of Tianjin Tasly Institute. The protocol was approved by the Animal Ethics Committee of Tianjin Tasly Institute.

The rats ($N = 4$) before lead injection were used as the control group and these rats after lead injection used as the test group. To induce ALP, the rats were intraperitoneally injected with $PbCl_2$ saline solution (5 mg lead per kg). For chemometric modeling, blood samples were collected from the control group and the test group 24 h post-injection. Blood samples were also collected from the test group 36 and 48 h post-injection for model validation. In addition, another control group ($N = 4$), namely a group with acute cadmium poisoning, was studied by intraperitoneally injecting the rats with $CdCl_2$ saline solution (5 mg cadmium per kg). The blood samples from this control group were collected 24 h post-injection. The blood samples were stored at about $-80$ °C for further treatment. Both $PbCl_2$ and $CdCl_2$ of analytical grade were obtained from local vendors.

## Plasma Sample Preparation

The blood sample was centrifuged at 3,000 rpm for 10 min, and a 10-μl aliquot of supernatant plasma was pipetted on the top of a piece of 1 × 1 cm aluminum foil. Each blood sample was used to prepare five replicate samples on aluminum foil. The foil was then placed in an oven set at 37°C for 2 h, and the obtained dry plasma film was subsequently used for FTIR measurement.

## FTIR Measurement

FTIR measurements were carried out on a Bruker Vertex 70 FTIR spectrometer (Ettlingen, Germany) equipped with a DLaTGS detector in attenuated total reflection (ATR) mode. 4 $cm^{-1}$ resolution and 32 scans were used for each measurement. A Pike Technologies MIRacle single-reflection ATR accessory (Madison, USA) with a diamond element was employed. When performing spectral acquisition, the plasma sample was pressed against the diamond crystal using a pressing device from Pike Technologies for a close contact. For each piece of aluminum foil with blood sample, at least seven FTIR spectra were taken by measuring signals at different locations on the foil.

## Spectral Pretreatment

The obtained FTIR spectra of the plasma samples were first screened to remove some error-based large deviation spectra. In ATR-FTIR mode, the contact between the sample and diamond crystal has a significant effect on the spectral quality, e.g., a poor contact will lead to poor quality FTIR spectra (abnormally low absorbance). These spectra need to be removed from the spectral dataset before chemometric analysis. Such spectral deviation is not due to the intrinsic deviation of one sample from its group (i.e., the control or test groups), but purely related to the spectral artifact caused by an improper contact between the sample and diamond crystal. These "abnormal" spectra could be easily identified visually with OPUS software and they

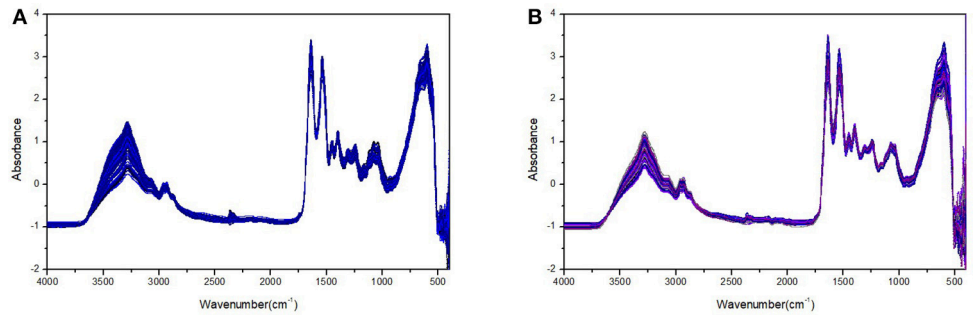



**FIGURE 1 |** Plasma FTIR spectra of the rat group without ALP **(A)** and with ALP **(B)** after spectral pretreatment such as smoothing, baseline correction, and vector normalization. The spectra with ALP were collected 24 h post-injection and there are a total of 139 spectra included in **(A)** and a total of 125 spectra included in **(B)**.

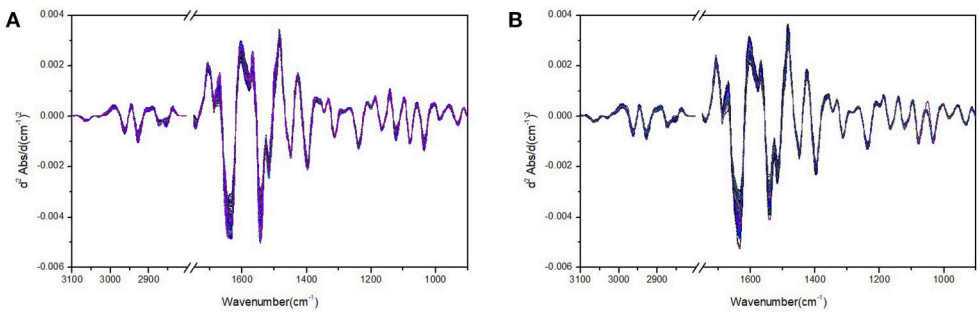



**FIGURE 2 |** Plasma FTIR second derivative spectra of the rat group without ALP **(A)** and with ALP **(B)** in the 3,100–2,800 and 1,750–900 $cm^{-1}$ spectral regions.

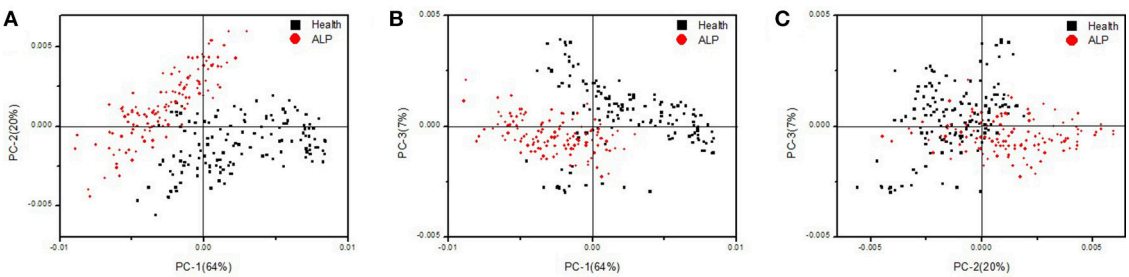



**FIGURE 3 |** Two-dimensional score plots of PC-1 vs. PC-2 **(A)**, PC-1 vs. PC-3 **(B)**, and PC-2 vs. PC-3 **(C)** obtained after PCA applied to the FTIR second derivative spectra of the rat groups without and with ALP.

were then removed from the spectral dataset manually. The remaining spectra were used for chemometric analysis after being subjected to spectral pre-treatment including smoothing, scattering correction, vector normalization and second derivative treatment with chemometric software.

## Chemometric Analysis

Chemometric analysis was performed using Unscrambler software (version 10.4) for PCA and partial least squares discriminant analysis (PLS-DA). In our study, we selected the data from the second derivative FTIR spectra in the regions of 3,100–2,800 and 1,750–900 $cm^{-1}$ for PCA. In addition, we also used 4-fold cross validation to test rat inter-individual variability on the spectra. The above-mentioned chemometric approach is relatively simple and sufficiently powerful to help differentiate the rat groups with and without ALP, spectroscopically.

# RESULTS AND DISCUSSION

**Figure 1** shows the plasma FTIR spectra of the rat groups without and with ALP after spectral pretreatment such as smoothing, baseline correction, and vector normalization. On the other hand, **Figure 2** shows the second derivative spectra of the plasma FTIR spectra presented in **Figure 1**. These second derivative spectra were the dataset used in the following chemometric analysis. The reason to have derivative treatment on the absorbance spectra in **Figure 1** is 2-fold. First, the second derivative treatment can further magnify the spectral changes and differences between the control and test groups. Second, the second derivative treatment can also eliminate possible interference of the baseline in chemometric analysis. In addition, in **Figure 2**, we have only included the spectral regions of 3,100–2,800 and 1,750–900 $cm^{-1}$ and removed the spectral region of 2,800–1,750 $cm^{-1}$ (as this region contains very limited spectral information). The 3,100–2,800 $cm^{-1}$ region corresponds to the C-H stretching absorptions; whereas the 1,750–900 $cm^{-1}$ corresponds to the protein amide I and amide II regions, and the fingerprint region. The displayed spectral regions in **Figure 2** contain most of the spectral information that is highly correlated to the ALP-induced biochemical changes in the plasma, thus making them suitable in our chemometric analysis.

As for the control and test group spectra datasets, we first used the most basic chemometric approach, PCA, to perform data analysis. We found the contribution rates of the first five principal components (namely PC-1, PC-2, PC-3, PC-4, and PC-5) are 64, 20, 7, 3, and 2%, respectively. The cumulative contribution rate of these five principal components reaches 96%, indicating that they can reflect most of the spectral variations and differences among the spectra of the control and test groups.

The two-dimensional score plots of PC-1 vs. PC-2, PC-1 vs. PC-3 and PC-2 vs. PC-3 were respectively shown in **Figures 3A–C**. Among the three score plots, we can clearly see that the two groups are well separated (**Figure 3A**) or they still have some significant overlaps (**Figures 3B**,**C**). **Figure 3A** gives the best discriminant result for the control and test groups. Our chemometric analysis study obviously demonstrates that with just some simple chemometric approaches such as PCA and PLS-DA, FTIR spectroscopy can be used to discriminate the rat groups with and without ALP.

For 4-fold cross validation on our data, each sample was used once as a test set while the remaining samples formed the training set. The results show that (i) there are significant differences between the test and control groups of plasma due to ALP and (ii) rat inter-individual variability has little influence on the spectral differences between the two groups. First, we analyzed the regions of 3,100–2,800 and 1,750–900 $cm^{-1}$ with PLS-DA. As displayed in **Figure 4**, the Y-variance plot shows that the line was basically leveled at PC7, and the more PCs could be overfitting; so seven PCs were selected for further analysis. **Figure 5** shows that PLS-DA could distinguish between health and ALP rats completely with seven PCs. However, the blue and red models of cross validation (CV) were not well matched. So, the fingerprint region of 1,750–900 $cm^{-1}$ was selected. As displayed in **Figure 6**, the Y-variance plot shows that seven PCs should be selected for further analysis. **Figure 7** shows not only that PLS-DA can distinguish between health and ALP rats completely with seven PCs, but also that the blue model fits well with the red CV model. In addition, the health and ALP groups in the red CV model are well separated by the 0.5 threshold line. In summary, the plasma spectra of health and ALP rats were distinctly different and inter-individual variability had no impact on the discrimination analysis of health and ALP rats.

The selectivity and robustness of our proposed PLS-DA model were also tested with additional controls to evaluate whether this model can give a correct discrimination when (i) when the rats suffer from another heavy metal poisoning and (ii) when the rats suffer from different extents of ALP. To address the first issue, we developed an acute cadmium poisoning rat model. Rats were injected with $CdCl_2$ solution to induce acute poisoning and the blood samples were collected 24 h post-injection. The plasma FTIR spectra and corresponding derivatives of this control group are presented in Figure S1 in the Supplementary Material. The

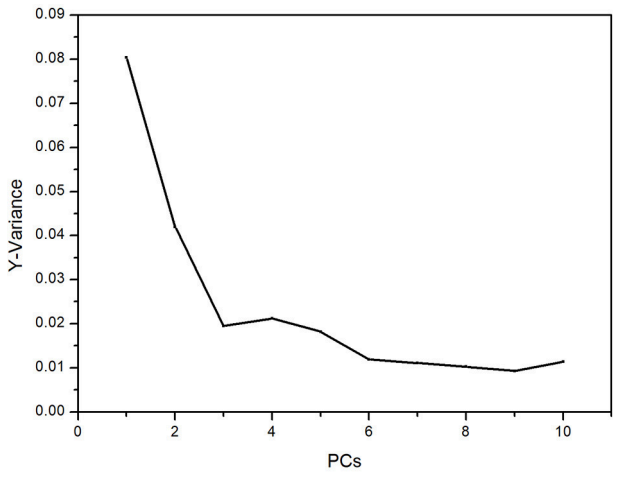



**FIGURE 4 |** PLS-DA Y-variance plot in the 3,100–2,800 and 1,750–900 $cm^{-1}$ spectral regions.

data with this control group were tested with our PLS-DA model. As we have mentioned above, the 0.5 value line is the threshold in the PLS-DA model in **Figure 7**. For data points above this line, the model predicts the rats are in ALP status; for data points below this line, the model predicts the rats are not in ALP status. As displayed in **Figure 8**, the predicted values for the rats suffering from acute cadmium poisoning are all below the 0.5 threshold, indicating that our PLS-DA model predicts that the rats suffering from cadmium poisoning are not in ALP status. This is a correct discrimination. To address the second issue, we performed a time-dependent study (up to 48 h post-injection) on the ALP rat model. The rats exposed to lead poisoning for different periods of time would suffer from lead poisoning to different extents. The plasma FTIR spectra and corresponding derivatives of this control group are presented in Figures S2, S3 in the Supplementary Material. We tested the 36 and 48 h data with our PLS-DA model. As we can see in **Figure 9**, the predicted values for these two control rat groups are all above the 0.5 threshold, indicating that these samples are in ALP status. This is a correct discrimination. These additional control experiments support the fact that our PLS-DA model is robust for ALP prediction.

Basically, some lead-induced biochemical changes in the plasma can be sensitively captured with chemometrics-based FTIR spectroscopy. To gain more insight into the biochemical changes induced by ALP in the plasma, the PLS-DA regression coefficient plot could be used to reflect corresponding spectral changes. As shown in **Figure 10**, this plot corresponds to the ALP-induced change in the composition and structure of the biochemical components in the plasma including biomacromolecular constitutes (such as proteins, DNAs and RNAs) as well as small molecular constitutes and metabolites

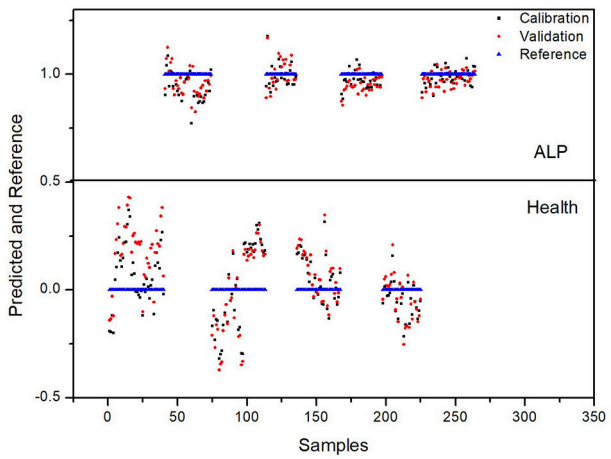



**FIGURE 5 |** PLS-DA predicted and reference plots in the 3,100–2,800 and 1,750–900 $cm^{-1}$ spectral regions.

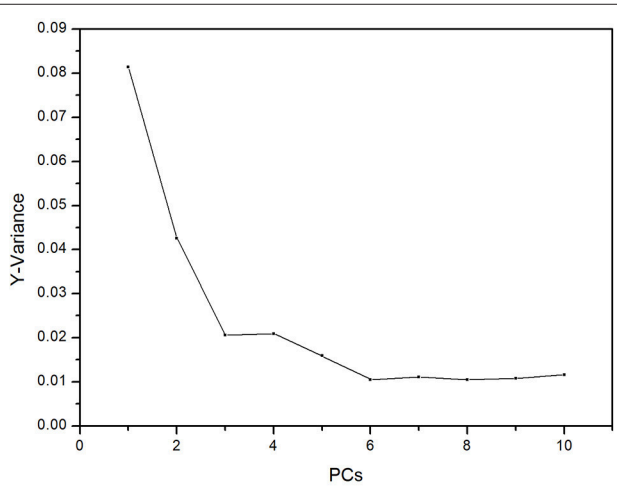



**FIGURE 6 |** PLS-DA Y-variance plot in the 1,750–900 $cm^{-1}$ spectral region.

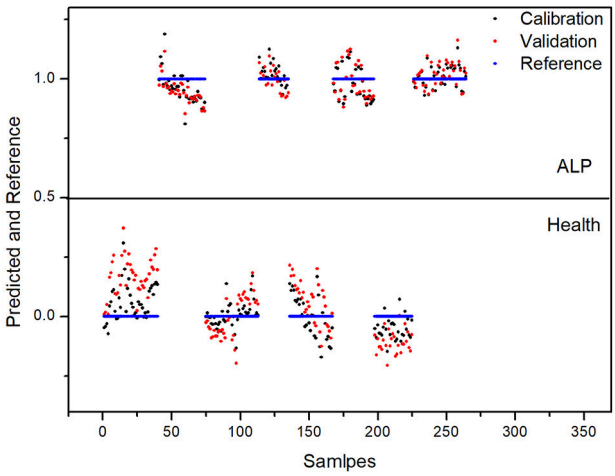



**FIGURE 7 |** PLS-DA predicted and reference plot in the 1,750–900 $cm^{-1}$ spectral region.

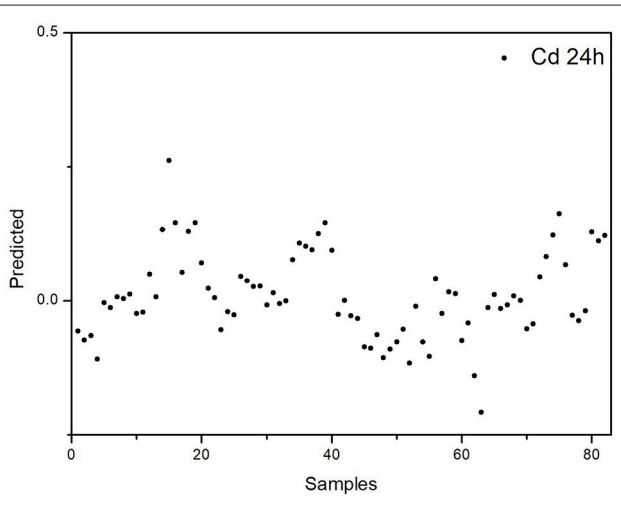



**FIGURE 8 |** Discrimination of the rats with acute cadmium poisoning using the proposed PLS-DA model for ALP.

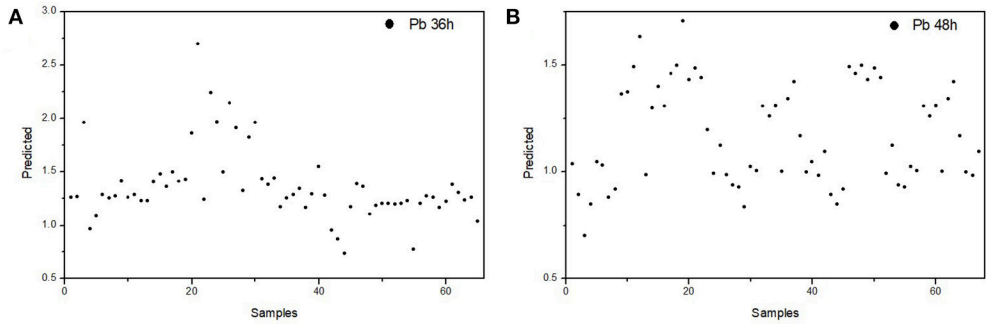



**FIGURE 9 |** Discrimination of the rats suffering from different extents of ALP using the proposed PLS-DA model: **(A)** 36h post-injection; **(B)** 48h post-injection.

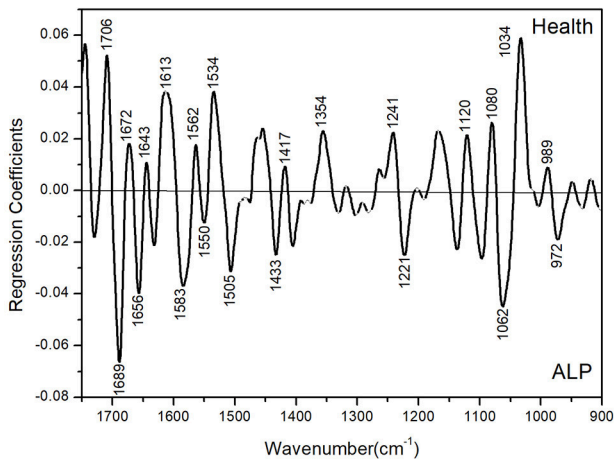



**FIGURE 10 |** PLS-DA regression coefficient plot in the 1,750–900 $cm^{-1}$ spectral region.

(such as lipids and carbohydrates). These plasma constitutes have characteristic vibrational absorptions in the PLS-DA regression coefficient plot. For example, through the spectral analysis of the 1,700–1,600 $cm^{-1}$ amide I region, we could obtain the information relevant to proteins; through the spectral analysis of the 1,300–1,000 $cm^{-1}$ region, we could obtain the information relevant to DNA and RNA. In addition, the intensity of the PLS-DA regression coefficient plot in different spectral regions could also provide information about the most prominent changes in the plasma. A summary is provided in **Table 1** for the spectral assignments for prominent peaks (either positive or negative) in the PLS-DA regression coefficient plot. They are based on the assignments in previous studies (Barth and Zscherp, 2002; Zandomeneghi et al., 2004; Zou et al., 2013; Staniszewska-Slezak et al., 2015). The peaks in the amide I (1,700–1,600 $cm^{-1}$) and amide II (around 1,550 $cm^{-1}$) correspond to absorptions of plasma proteins. In this region, we observed several prominent peaks in the PLS-DA regression coefficient plot including the amide I and amide II peaks at 1,706, 1,689, 1,672, 1,656, 1,643, 1,613, 1,550, and 1,534 $cm^{-1}$. This observation in the PLS-DA regression coefficient plot suggests that ALP induced significant compositional and structural changes of the proteins in the plasma of the ALP rat model. Such changes may be due to the direct coordination effect of lead ion with protein or be due to the perturbation of lead ion on the biosynthesis of proteins in the rat. In addition, lead ion may interact (or coordinate) with the side chains of some amino acids (such as tryptophan, histidine, aspartic acid, and glutamic acid) or affect the biosynthesis of these amino acids. Such interactions or perturbations are suggested by the observation of the peaks at 1,505, 1,354, and 1,241 $cm^{-1}$ (corresponding to the side chain of tryptophan), at 1,583 and 1,433 $cm^{-1}$ (corresponding to the side chain of histidine) and at 1,417 $cm^{-1}$ (corresponding to the side chains of aspartic acid and glutamic acid). The PLS-DA regression coefficient plot also suggests that the nucleic acid, DNA and RNA changes in the plasma as the peaks at 1,221, 1,120, 1,080, and 1,062 $cm^{-1}$ are observed in the regression coefficient. These peaks correspond to the $PO_2^-$ and C-O absorption of DNA and RNA. At last, the peaks at 1,034 $cm^{-1}$ (which may be related to the metabolism of glucose and polysaccharides) and at 989 and 972 $cm^{-1}$ (which corresponds to the phosphorylation modification of proteins) are also observed in the regression coefficient plot. In summary, on the one hand, the PLS-DA regression coefficient plot suggests a very complex biochemical changes that occurred in the body of the lead-poisoned rats; one the other hand, ALP-induced protein changes seem to be the most important cause for the rat poisoning. This finding further implies that the spectral changes corresponding to the biochemical changes of proteins may be used as potential spectral biomarkers for the diagnostics of ALP.

## CONCLUSION

In this exploratory study, we have demonstrated that FTIR spectroscopy empowered with PCA and PLS-DA analysis can capture ALP-induced biochemical changes in the plasma spectroscopically and is capable of differentiating the rats with and without suffering from ALP. Furthermore, the revealed FTIR spectral changes, in particular, corresponding to the biochemical changes of proteins, may be used as potential spectral biomarkers for the diagnostics of lead poisoning. Our method has sufficient discriminant ability and the potential to be employed as a blood-based objective, convenient, and non-destructive diagnostic tool

**TABLE 1 |** Spectral assignment for the observed peaks in the PLS-DA regression coefficient plot.

| Peak position ($cm^{-1}$) | Spectral assignment |
|---|---|
| 1,706 | Protein amide I |
| 1,689 | Protein amide I |
| 1,672 | Protein amide I |
| 1,656 | Protein amide I |
| 1,643 | Protein amide I |
| 1,613 | Protein amide I |
| 1,583 | C=C vibration of histidine |
| 1,562 | Protein amide II |
| 1,550 | Protein amide II |
| 1,534 | Protein amide II |
| 1,505 | Indole vibration of tryptophan |
| 1,433 | C-N vibration of histidine |
| 1,417 | C-C, C-H, and N-H vibrations of tryptophan |
| 1354 | Indole vibration of tryptophan |
| 1,241 | C-H and C-C vibrations of tryptophan |
| 1,221 | $PO_2^-$ antisymmetric stretch of nucleic acids, DNA, and RNA |
| 1,120 | C-O stretch of DNA and RNA |
| 1,080 | $PO_2^-$ vibrations of nucleic acids, phospholipids, and saccharids |
| 1,062 | $PO_2^-$ symmetric stretch of nucleic acids, DNA, and RNA |
| 1,034 | C-O-H bend of glucose and polysaccharide |
| 989 | Protein phosphorylation |
| 972 | Protein phosphorylation |

for lead poisoning. To the best of our knowledge, this work is the first application of chemometrics-based FTIR spectroscopy in the diagnostics of lead poisoning. We hope the chemometrics-based FTIR spectroscopy can evolve into an objective, convenient, cost-effective and non-destructive disease diagnostics tool in the future.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of institutional guidelines of the Animal Ethics Committee of Tianjin Tasly Institute. The protocol was approved by the Animal Ethics Committee of Tianjin Tasly Institute.

## AUTHOR CONTRIBUTIONS

WT and GM designed the project. WT, DW, HF, LY, and GM conducted the experiments and analysed the data. GM, WT, and HF wrote the manuscript.

## ACKNOWLEDGMENTS

We gratefully acknowledge the financial support from the National Natural Science Foundation of China (No. 21075027), the Natural Science Foundation of Hebei Province (Nos. B2011201082 and B2016201034), Juren plan, and Program for Changjiang Scholars and Innovative Research Team in University (No. IRT_15R16). GM thanks Xiangke Chen for helpful discussions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00261/full#supplementary-material

## REFERENCES

Baker, M. J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H. J., Dorling, K. M., et al. (2014). Using fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* 9, 1771–1791. doi: 10.1038/nprot.2014.110

Barth, A., and Zscherp, C. (2002). What vibrations tell us about proteins. *Q. Rev. Biophys.* 35, 369–430. doi: 10.1017/s0033583502003815

Brodkin, E., Copes, R., Mattman, A., Kennedy, J., Kling, R., and Yassi, A. (2007). Lead and mercury exposures: interpretation and action. *Can. Med. Assoc. J.* 176, 59–63. doi: 10.1503/cmaj.060790

De Bruyne, S., Speeckaert, M. M., and Delanghe, J. R. (2018). Applications of mid-infrared spectroscopy in the clinical laboratory setting. *Crit. Rev. Clin. Lab. Sci.* 55, 1–20. doi: 10.1080/10408363.2017.1414142

Deleris, G., and Petibois, C. (2003). Applications of FT-IR spectrometry to plasma contents analysis and monitoring. *Vib. Spectrosc.* 32, 129–136. doi: 10.1016/s0924-2031(03)00053-5

Depciuch, J., Kaznowska, E., Koziorowska, A., and Cebulski, J. (2017). Verification of the effectiveness of the Fourier transform infrared spectroscopy computational model for colorectal cancer. *J. Pharm. Biomed. Anal.* 145, 611–615. doi: 10.1016/j.jpba.2017.07.026

Ellis, D. I., and Goodacre, R. (2006). Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *Analyst* 131, 875–885. doi: 10.1039/b602376m

Elmi, F., Movaghar, A. F., Elmi, M. M., Alinezhad, H., and Nikbakhsh, N. (2017). Application of FT-IR spectroscopy on breast cancer serum analysis. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 187, 87–91. doi: 10.1016/j.saa.2017.06.021

Flegal, A. R., and Smith, D. R. (1992a). Current needs for increased accuracy and precision in measurements of low levels of lead in blood. *Environ. Res.* 58, 125–133. doi: 10.1016/s0013-9351(05)80209-9

Flegal, A. R., and Smith, D. R. (1992b). Lead levels in preindustrial humans. *N. Engl. J. Med.* 326, 1293–1294.

Gajjar, K., Trevisan, J., Owens, G., Keating, P. J., Wood, N. J., Stringfellow, H. F., et al. (2013). Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst* 138, 3917–3926. doi: 10.1039/c3an36654e

Gasper, R., Dewelle, J., Kiss, R., Mijatovic, T., and Goormaghtigh, E. (2009). IR spectroscopy as a new tool for evidencing antitumor drug signatures. *Biochim. Biophys. Acta Biomembr.* 1788, 1263–1270. doi: 10.1016/j.bbamem.2009.02.016

Ghimire, H., Venkataramani, M., Bian, Z., Liu, Y., and Perera, A. G. U. (2017). ATR-FTIR spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples. *Sci. Rep.* 7:16993. doi: 10.1038/s41598-017-17027-4

Goldstein, G. W. (1992). Neurological concepts of lead poisoning in children. *Pediatr. Ann.* 21, 384–388.

Guo, F., Zhu, Y., Chen, C., Wang, S., and Liang, S. (2017). Construction of different calibration models by FTIR/ATR spectra and their application in screening of phenylketonuria. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 177, 33–40. doi: 10.1016/j.saa.2017.01.020

Hunter, D. (1978). *The Disease of Occupations.* Sevenoaks: Hodder and Stoughton.

Kazantzis, G. (1989). "Lead: ancient metal—modern menace?" in *Lead Exposure and Child Development: an International Assessment*, eds M. A. Smith, L. D. Grant, and A. L. Soris (Lancaster: MTP Press), 119–128.

Krafft, C., Sobottka, S. B., Geiger, K. D., Schackert, G., and Salzer, R. (2007). Classification of malignant gliomas by infrared spectroscopic imaging and linear discriminant analysis. *Anal. Bioanal. Chem.* 387, 1669–1677. doi: 10.1007/s00216-006-0892-5

Krafft, C., Steiner, G., Beleites, C., and Salzer, R. (2009). Disease recognition by infrared and Raman spectroscopy. *J. Biophotonics* 2, 13–28. doi: 10.1002/jbio.200810024

Le Corvec, M., Jezequel, C., Monbet, V., Fatih, N., Charpentier, F., Tariel, H., et al. (2017). Mid-infrared spectroscopy of serum, a promising non-invasive method to assess prognosis in patients with ascites and cirrhosis. *PLoS ONE* 12:e0185997. doi: 10.1371/journal.pone.0185997

Li, Z., Lv, H., Li, T., Si, G., Wang, Q., Lv, J., et al. (2017). Reagent-free simultaneous determination of glucose and cholesterol in whole blood by FTIR-ATR. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 178, 192–197. doi: 10.1016/j.saa.2017.02.002

Liu, H., Su, Q., Sheng, D., Zheng, W., and Wang, X. (2017). Comparison of red blood cells from gastric cancer patients and healthy persons using FTIR spectroscopy. *J. Mol. Struct.* 1130, 33–37. doi: 10.1016/j.molstruc.2016.10.019

Mitchell, A. L., Gajjar, K. B., Theophilou, G., Martin, F. L., and Martin-Hirsch, P. L. (2014). Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting. *J. Biophotonics* 7, 153–165. doi: 10.1002/jbio.201400018

Ollesch, J., Heinze, M., Heise, H. M., Behrens, T., Brüning, T., and Gerwert, K. (2014). It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy. *J. Biophotonics* 7, 210–221. doi: 10.1002/jbio.201300163

Paraskevaidi, M., Morais, C. L. M., Lima, K. M. G., Snowden, J. S., Saxon, J. A., Richardson, A. M. T., et al. (2017). Differential diagnosis of Alzheimer's disease using spectrochemical analysis of blood. *Proc. Natl. Acad. Sci. U S A.* 114, E7929–E7938. doi: 10.1073/pnas.1701517114

Patrick, L. (2006). Lead toxicity, a review of the literature. Part 1: exposure, evaluation, and treatment. *Altern. Med. Rev.* 11, 2–22.

Rai, V., Mukherjee, R., Routray, A., Ghosh, A. K., Roy, S., Ghosh, B. P., et al. (2018). Serum-based diagnostic prediction of oral submucous fibrosis using FTIR spectrometry. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 189, 322–329. doi: 10.1016/j.saa.2017.08.018

Roy, S., Perez-Guaita, D., Andrew, D.W., Richards, J.S., McNaughton, D., Heraud, P., et al. (2017). Simultaneous ATR-FTIR based determination of malaria parasitemia, glucose and urea in whole blood dried onto a glass slide. *Anal. Chem.* 89, 5238–5245. doi: 10.1021/acs.analchem.6b04578

Sarkar, A., Sengupta, S., Mukherjee, A., and Chatterjee, J. (2017). Fourier transform infra-red spectroscopic signatures for lung cells' epithelial mesenchymal transition: a preliminary report. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 173, 809–816. doi: 10.1016/j.saa.2016.10.019

Sheng, D., Xu, F., Yu, Q., Fang, T., Xia, J., Li, S., et al. (2015). A study of structural differences between liver cancer cells and normal liver cells using FTIR spectroscopy. *J. Mol. Struct.* 1099, 18–23. doi: 10.1016/j.molstruc.2015.05.054

Smith, M. A. (1984). "Lead in history," in *The Lead Debate: the Environmental Toxicology and Child Health*, eds R. Lansdown and W. Yule (London: Croom Helm), 7–24.

Staniszewska-Slezak, E., Fedorowicz, A., Kramkowski, K., Leszczynska, A., Chlopicki, S., Baranska, M., et al. (2015). Plasma biomarkers of pulmonary hypertension identified by Fourier transform infrared spectroscopy and principal component analysis. *Analyst* 140, 2273–2279. doi: 10.1039/c4an01864h

Titus, J., Viennois, E., Merlin, D., and Perera, A. G. U. (2017). Minimally invasive screening for colitis using attenuated total internal reflectance fourier transform infrared spectroscopy. *J. Biophotonics* 10, 465–472. doi: 10.1002/jbio.201600041

Tong, S. (1998). Lead exposure and cognitive development: persistence and a dynamic pattern. *J. Paediatr. Child Health* 34, 114–118. doi: 10.1046/j.1440-1754.1998.00187.x

Tong, S., von Schirnding, Y. E., and Prapamontol, T. (2000). Environmental lead exposure: a public health problem of global dimensions. *Bull. World Health Organ.* 78, 1068–1077.

Winder, C. (1984). *The Developmental Neurotoxicity of Lead.* Lancaster: MTP Press.

Zandomeneghi, G., Krebs, M.R., McCammon, M.G., and Fändrich, M. (2004). FTIR reveals structural differences between native beta-sheet proteins and amyloid fibrils. *Protein Sci.* 13, 3314–3321. doi: 10.1110/ps.04102404

Zou, Y., Li, Y., Hao, W., Hu, X., and Ma, G. (2013). Parallel beta-sheet fibril and antiparallel beta-sheet oligomer: new insights into amyloid formation of hen egg white lysozyme under heat and acidic condition from FTIR spectroscopy. *J. Phys. Chem. B* 117, 4003–4013. doi: 10.1021/jp4003559

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 06 July 2018
doi: 10.3389/fchem.2018.00262

# Pharmaceutical Analysis Model Robustness From Bagging-PLS and PLS Using Systematic Tracking Mapping

*Na Zhao*[1], *Lijuan Ma*[2,3], *Xingguo Huang*[2,3], *Xiaona Liu*[4], *Yanjiang Qiao*[1,2,3]* *and Zhisheng Wu*[1,2,3]*

[1] Key Laboratory of Xinjiang Phytomedicine Resources and Utilization, Ministry of Education, School of Pharmacy, Shihezi University, Shihezi, China, [2] Beijing University of Chinese Medicine, Beijing, China, [3] Pharmaceutical Engineering and New Drug Development of TCM of Ministry of Education, Beijing, China, [4] School of Integrated Traditional Chinese and Western Medicine, Binzhou Medical University, Yantai, China

Our work proved that processing trajectory could effectively obtain a more reliable and robust quantitative model compared with the step-by-step optimization method. The use of systematic tracking was investigated as a tool to optimize modeling parameters including calibration method, spectral pretreatment and variable selection latent factors. The variable was selected by interval partial least-squares (iPLS), backward interval partial least-square (BiPLS) and synergy interval partial least-squares (SiPLS). The models were established by Partial least squares (PLS) and Bagging-PLS. The model performance was assessed by using the root mean square errors of validation (RMSEP) and the ratio of standard error of prediction to standard deviation (RPD). The proposed procedure was used to develop the models for near infrared (NIR) datasets of active pharmaceutical ingredients in tablets and chlorogenic acid of *Lonicera japonica* solution in ethanol precipitation process. The results demonstrated the processing trajectory has great advantages and feasibility in the development and optimization of multivariate calibration models as well as the effectiveness of bagging model and variable selection to improve prediction accuracy and robustness.

**Keywords: multivariate calibration, near infrared spectroscopy, processing trajectory, Bagging-PLS, variable selection**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Huawen Wu,*
*BaySpec, Inc., United States*
*Francesco Crea,*
*Università degli Studi di Messina, Italy*

***Correspondence:**
*Yanjiang Qiao*
*yjqiao@263.net*
*Zhisheng Wu*
*wzs@bucm.edu.cn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *28 November 2017*
***Accepted:*** *12 June 2018*
***Published:*** *06 July 2018*

***Citation:***
*Zhao N, Ma L, Huang X, Liu X, Qiao Y and Wu Z (2018) Pharmaceutical Analysis Model Robustness From Bagging-PLS and PLS Using Systematic Tracking Mapping. Front. Chem. 6:262. doi: 10.3389/fchem.2018.00262*

## INTRODUCTION

Multivariate calibration is the process of relating the measured response to the analyte amounts, concentrations, or other measured values of physical or chemical properties. Partial least squares (PLS) regression is the most effective and commonly used regression techniques in multivariate calibration because of its calibration model quality and ease of implementation. The statistical results show that approximately 20,000 published papers reports used PLS models from 2005 to 2017. The PLS technique has been effectively applied to different fields, especially in pharmaceutical analysis.

Kachrimanis et al. developed a fast and precise method using FT-Raman spectroscopy alongside with PLS for the quantitation of monoclinic and orthorhombic paracetamol in powder mixtures (Kachrimanis et al., 2007). Yu et al. established a PLS model using near infrared spectroscopy (NIR) and gas chromatography data to determine *l*-borneol in Blumea balsamifera (Ai-na-xiang) samples

(Yu et al., 2017). Sarkhosh et al. developed a PLS model of redox potential with genetic algorithms selecting pixels in multivariate image analysis for a quantitative structure-activity relationships (QSAR) study of trypanocidal activity for quinone compounds (Sarkhosh et al., 2014). Üstün et al. built a fast quantification method combining $^{1}$H NMR spectroscopy with PLS to determine the chondroitin sulfate and dermatan sulfate in danaparoid sodium (Üstün et al., 2011). Wu et al. used NIR as a process analytical technology and developed the PLS model of 11 amino acids to monitor their concentration change during hydrolysis process of Cornu Bubali (Wu et al., 2013b).

The successful application of PLS depends on the development and validation of multivariable models. Recently, the multivariate data needs a more suitable method to establish a robust and reliable PLS model. However, many parameters need to be optimized for a quantitative PLS model, which include spectral pretreatment, variable selection, calibration methods, etc. To improve model performance, the pretreatments are used to reduce the undesirable variations effects from instrument, environment, sample preparation protocol, etc. (Faber, 1999; Blanco et al., 2007; Fernández-Cabanás et al., 2007; Lim et al., 2016).

Besides, variable selection in modeling is also an important step to identify informative features and/or remove uninformative variables for better prediction performance and model complexity reduction. Recently, based on the PLS algorithm, some variable selection methods have been developed including interval partial least-squares (iPLS) (Saudland et al., 2000), backward interval partial least-square (BiPLS) (Leardi and Nørgaard, 2004) and synergy interval partial least-squares (SiPLS) (Munck et al., 2001), etc. Many studies have confirmed the efficiency of these variable selection methods for improving model performance (Chen et al., 2008; Di et al., 2010; Wu et al., 2013a; Mahanty et al., 2016).

In addition, a single model is often not robust because of the change of calibration data and model parameters. An alternative effective approach to improve model robustness is ensemble modeling that establishes multiple models and combines their predictions into a single value. Bagging-PLS is one of most important ensemble modeling techniques. About 60 papers were published on the use of Bagging-PLS model in the period 2005–2017. Galvão et al. used bagging strategies in conjunction with Multiple Linear Regression (MLR) and PLS to develop the multivariate calibration models for four diesel quality parameters, showing that the prediction accuracy was improved by subagging procedure (Galvão et al., 2006). Pan et al. combined ensemble method of Bagging with PLS to detect naringin, hesperidin and neohesperidin in pilot-scale extraction process of *Fructus aurantii* with online NIR sensors (Pan et al., 2015).

Most of the published works dealing with PLS model used a univariate to optimize these modeling parameters step by step according to the root-mean-square error. The number of modeling paths of this method was limited and the results were often not the global optimal. Then, we proposed processing trajectory that can provide a systematic way to optimize parameters in a quantitative model (Zhao et al., 2015).

Based on the above considerations, we extend the optimization of spectral pretreatment, latent factors and variable selection using tracking procedure to spectral pretreatment, latent factors, variable selection and calibration method. The methods of variable selection included iPLS, BiPLS, and SiPLS. The models were established by using PLS and Bagging-PLS. The model performance was assessed using the root mean square errors of validation (RMSEP) and the ratio of standard error of prediction to standard deviation (RPD) (Esbensen et al., 2014; Williams et al., 2014). Two diferent NIR spectral datasets (one standard and one open source) were analyzed. The proposed procedure was used to predict active pharmaceutical ingredients (API) in tablets and chlorogenic acid of *Lonicera japonica* solution in ethanol precipitation process.

# DATASETS AND ANALYSIS

## Datasets

### Tablet

The NIR transmittance spectra of a pharmaceutical tablet were described in Dyrby et al. (2002) and publicly available at http://www.models.life.ku.dk/Tablets. This tablet dataset consists of 310 samples measured in the range of 7,000–10,500 cm$^{-1}$ with a

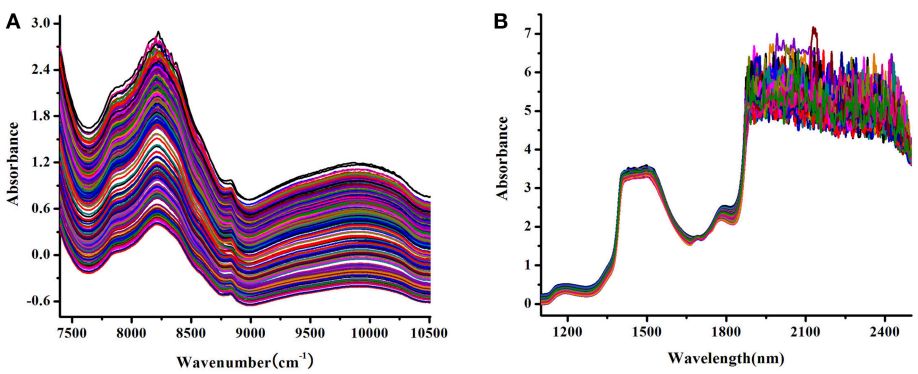



**FIGURE 1 |** Raw NIR spectra of tablet samples **(A)** and *Lonicera japonica* solution in ethanol precipitation process **(B)**.

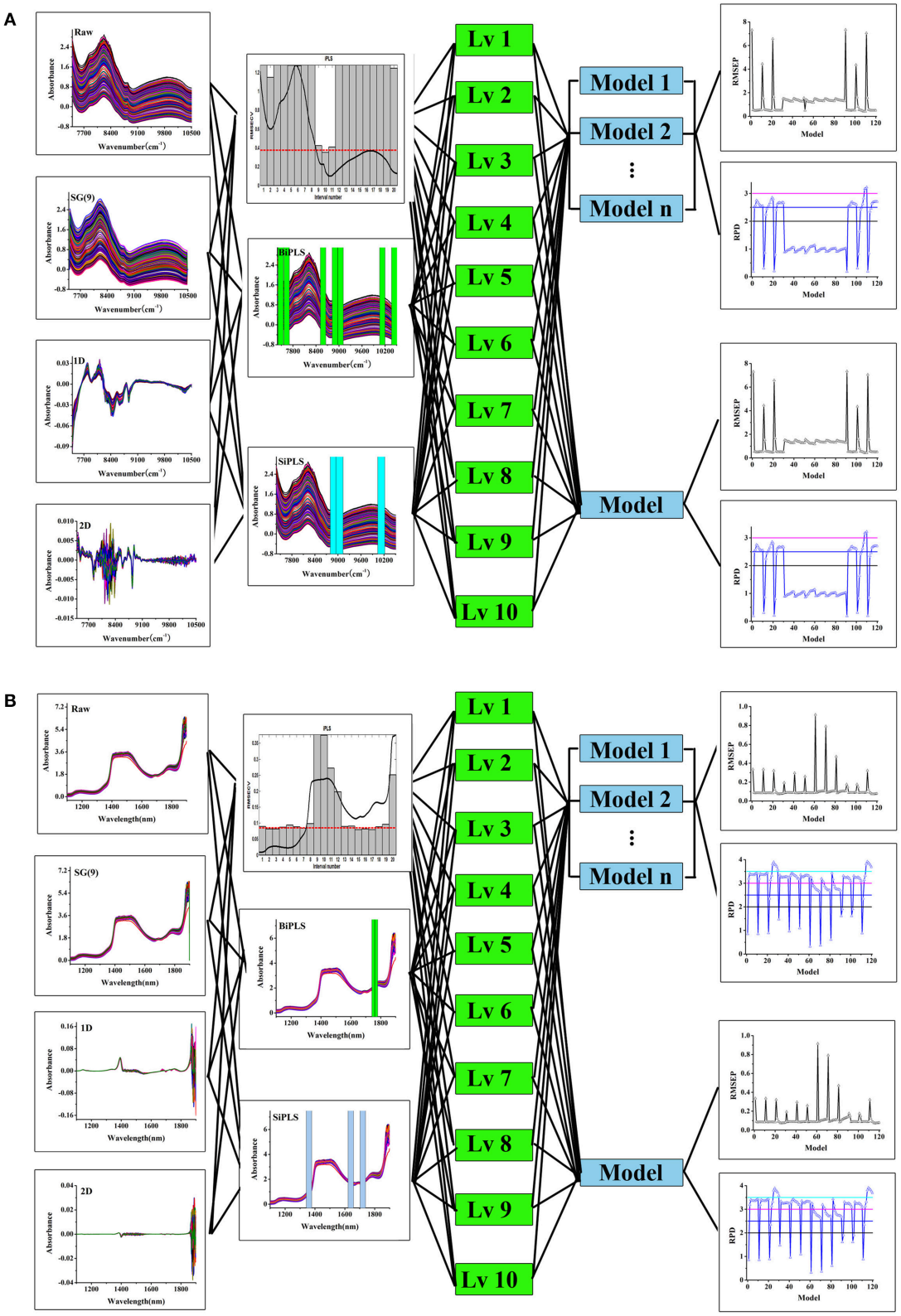



**FIGURE 2 |** The processing trajectory and assessment of PLS and Bagging-PLS model, tablet samples **(A)** and *Lonicera japonica* solution in ethanol precipitation process **(B)**.

resolution of 16 $cm^{-1}$ i.e., a total number of 404 variables per sample. The objective of the analysis was to predict the API content of the tablet. The content of API in the tablets (% w/w) was assayed by high performance liquid chromatography (HPLC). The tablet dataset was supplied in Data Sheet 1. This dataset was divided into two groups: 207 and 103 samples for training and validation with Kennard-Stone (KS) algorithm, respectively.

### *Lonicera japonica*

The NIR spectral dataset of *Lonicera japonica* has been reported previously (Wu et al., 2012). The data consisted of 216 samples with 2,800 variables in the range of 1,100–2,500 nm that measured on an XDS rapid liquid analyzer with VISION software in the transmission mode (Foss NIR Systems, Silver Spring, MD, USA). NIR spectra of *Lonicera japonica* solution obtained from ethanol precipitation process, were measured to estimate chlorogenic acid content. HPLC was used as the reference method for chlorogenic acid determination as recommended by the Chinese Pharmacopoeia (CHP, 2010 Edition) for *Lonicera japonica* monograph. The dataset of *Lonicera japonica* was supplied in Data Sheet 2. In this study, the training data consisted of 144 samples and the remaining 72 samples were used for validation.

### Multivariate Data Analyses

The spectral pretreatment of data was performed using chemometric tool in this study (SIMCA P + 11.5, Umetrics, Sweden). Data analysis was conducted using Unscrambler 9.7 software package (Camo Software AS, Norway) and Matlab version 7.0 (MathWorks Inc., USA). Some of the algorithms were developed by Norgaard et al., readily downloadable from http://www.models.life.ku.dk/iToolbox.

### Multivariate Calibration

A procedure for the development and optimization of multivariate calibration models using processing trajectory is summarized in **Figure 2**. The rationale behind this approach is that there was more than one path to obtain good model with different parameter combinations. Thus, the procedure was used to track and evaluate modeling processes with different parameters including spectral pretreatments, variable selections, latent factors, and calibration methods. The evaluation indexes of model includes RMSEP and RPD.

## RESULT AND DISCUSSION

## Raw Spectra

The raw NIR spectra of the tablet and *Lonicera japonica* solution were shown in **Figure 1**, which represent their characteristic peak locations regarding the active substance in each spectral dataset. In the NIR transmittance spectra of tablet (**Figure 1A**), there were several broad peaks located at around 10,000, 8,830, 8,200, and 7,840 $cm^{-1}$, which originated from several components in the corresponding drug tablet. In addition, there were large fluctuations in the combined region of fundamental vibrations in the raw spectra of *Lonicera japonica* solution. Therefore, the spectral region of 1,100–1,900 nm was selected.

## Processing Trajectory of PLS Model

The modeling procedure using processing trajectory was showed in **Figure 2**. Taking the tablet dataset as an example, the data set were split in to calibration and validation sets and the

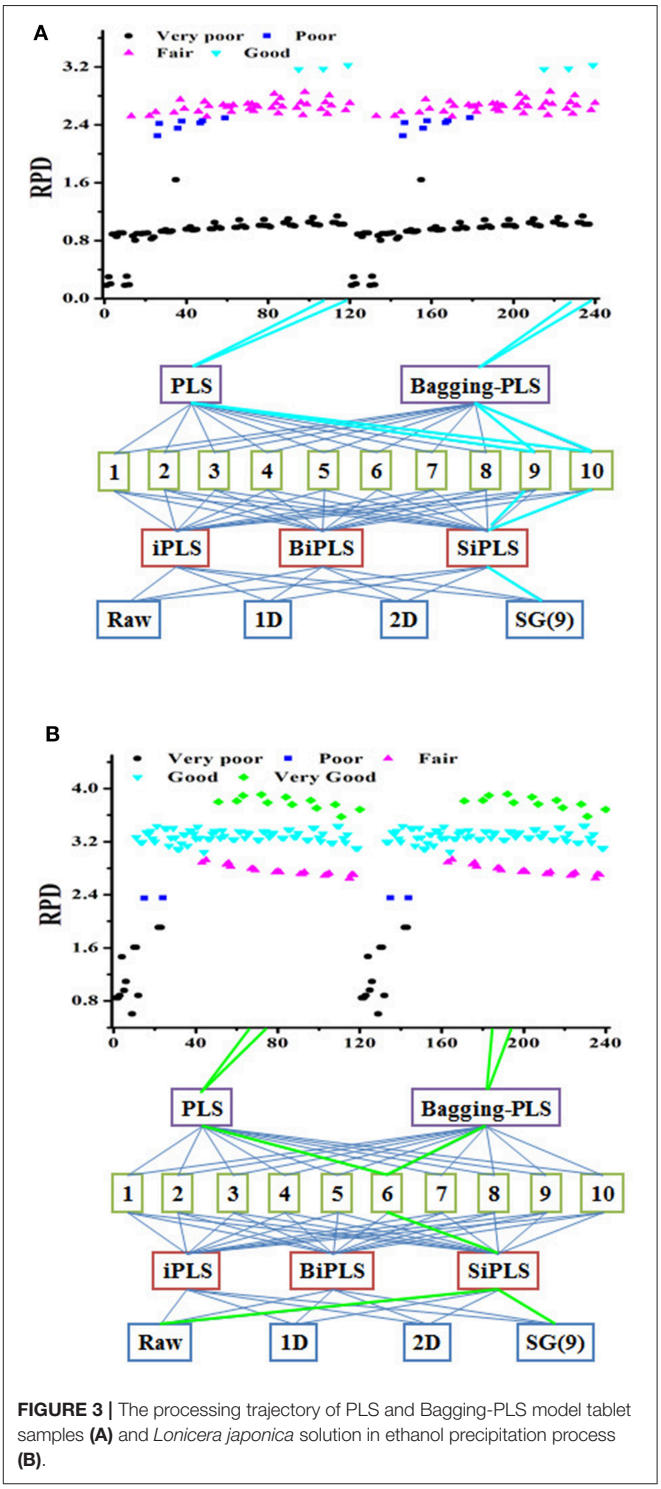



**FIGURE 3 |** The processing trajectory of PLS and Bagging-PLS model tablet samples **(A)** and *Lonicera japonica* solution in ethanol precipitation process **(B)**.

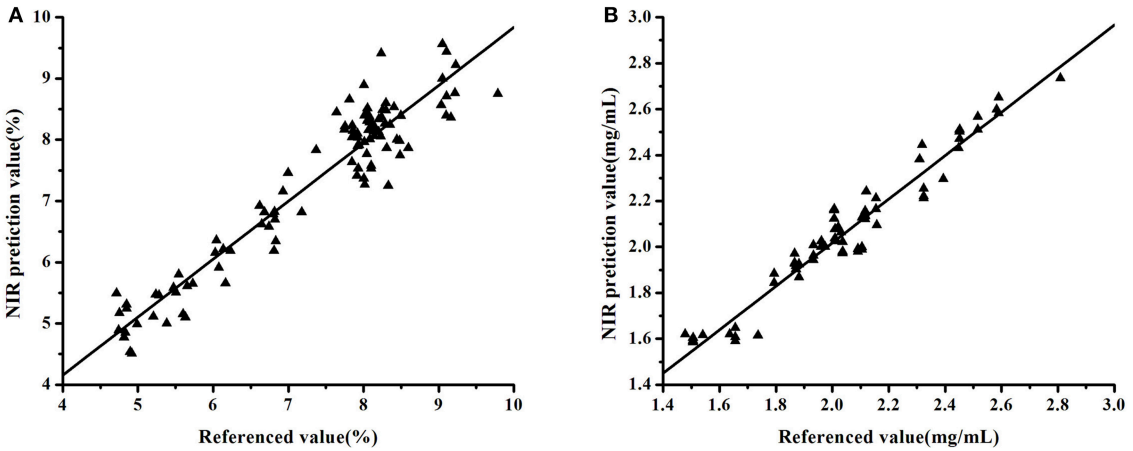



**FIGURE 4 |** Correlation between the prediction and reference values of the datasets: tablet samples **(A)** and *Lonicera japonica* solution in ethanol precipitation process **(B)**.

spectra were preprocessed using different methods including first derivative (1st), second derivative (2nd) and Savitzky-Golay smoothing with 9 points [SG(9)]. The iPLS, BiPLS and SiPLS were then used to select variables. Finally, the PLS and Bagging-PLS models were developed with latent factors from 1 to 10. Both RPD and RMSEP were calculated to evaluate the model. **Figure 2** showed different modeling paths and model results. The parameters for PLS and Bagging-PLS models of API in tablet and chlorogenic acid of *Lonicera japonica* solution were shown in Tables S1, S2.

The RPD and RMSEP had similar trends in PLS and Bagging-PLS models. In **Figure 2A**, the RMSEP decreased with increasing latent factor coupled with different pretreatment methods and variables selections. The RPD also increased with an increase of small latent factors. However, when the latent variable was greater than a certain value, the RPD became smaller. Variances in RMSEP and RPD indexes were not obvious when using 1st and 2nd derivative preprocessed spectra. Other pretreatment methods were superior to 1st and 2nd derivative processing. The model for *Lonicera japonica* dataset is shown in **Figure 2B**. Similar results were found for the tablet dataset. The model results of other pretreatment methods were also better than 2nd derivative processing.

Moreover, this finding indicates that more than one modeling path could ensure a successful model. Data obtained from different modeling paths and model classification were shown in **Figure 3**. There were six good models with RPD between 3 and 3.5 (**Figure 3A**), and some very good model paths with RPD values greater than 3.5 (**Figure 3B**). In the previous modeling process routine, the parameters were optimized one at a time according to the resultant prediction accuracy. This was a poor approach to path modeling vs. step-by-step parameter optimization (Table S3). The optimal parameters of the API model obtained step-by-step optimized were the raw spectra and iPLS-selecting variable under 3 latent factors. The model performance was fair. However, the result of processing trajectory showed that six good models could be obtained by combination of SG(9) pretreatment and BiPLS-selecting variables.

## Development and Validation of Calibration Models

The best nonsystematic parameter combination for the chlorogenic acid Bagging-PLS model was raw spectra and iPLS or BiPLS variables selection under 2 latent factors. The model performance was good. However, there were 24 very good models with different systematic parameter combinations in the result of processing trajectory. The best parameter combination of the chlorogenic acid model was that the model was developed by Bagging-PLS with SG(9) spectral pretreatment and SiPLS-selecting variables under 6 factors. It demonstrated that the model obtained through the processing trajectory was better than that step-by-step optimized. It means that the optimal systematic model parameter combination can be obtained via the processing trajectory and bagging ensemble modeling techniques, and variable assignment could improve prediction accuracy and robustness.

The model validity was evaluated in terms of RMSEP and RPD values. Taking the tablet dataset as an example, **Figure 2A** showed that the model established using Bagging-PLS with SG(9) pretreatment and BiPLS-selecting variables under 10 latent factors had the best performance. The RMSEP and RPD values of the validation set were 0.4126% and 3.2234, respectively. In contrast, the RMSEP and RPD of the model step-by-step optimized were 0.5164% and 2.5755, respectively. These results also showed that the model developed with Bagging-PLS had a good predictive performance. Similarly, the model of *Lonicera japonica* solution was developed using Bagging-PLS with SG(9) spectral pretreatment and SiPLS-selecting variables under 6 latent factors. The RMSEP and RPD were 0.0728 mg/mL and 3.9166, respectively. The RMSEP and RPD of the model step-by-step optimized were 0.0891% and 3.1966, respectively. **Figure 4** presents the data obtained with Bagging-PLS models using the two datasets. The prediction values reasonably agreed with

HPLC results. The parameters indicated that NIRS could be used for the determination of API in tablets and chlorogenic acid of *Lonicera japonica* solution in ethanol precipitation process.

## CONCLUSION

We proposed processing trajectory to optimize the parameters of multivariate calibration such as spectral pretreatment, latent factors, variable selection and calibration methods. The models were developed using PLS and Bagging-PLS with different spectral pretreatments and variable selection methods under different latent factors. The chemometric indicators (RMSEP and RPD) were used to evaluated the model. The different PLS and Bagging-PLS models were used to quantify the API in tablets and chlorogenic acid of *Lonicera japonica* solution in ethanol precipitation process. The result illustrated that the processing trajectory has great advantages and feasibility in the development and optimization of multivariate calibration models and the effectiveness of bagging model and variable selection to improve prediction accuracy and robustness.

In conclusion, the application of processing trajectory for model optimization shows excellent results to develop a reliable and robust model. The proposed should be translated into an algorithm to be integrated into PLS software, helping to obtain better models.

## AUTHOR CONTRIBUTIONS

YQ and ZW conceived and designed the study. NZ performed the experiment with the help of LM, XH, and XL. NZ and ZW wrote the manuscript. All authors read and approved the final manuscript.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (81773914), Beijing Nova Program of China (xx2016050), Science Fund for Distinguished Young Scholars in BUCM (2015-JYB-XYQ-003) and Fund for young teachers in BUCM (2016-JYB-JSMS-061).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00262/full#supplementary-material

## REFERENCES

Blanco, M., Castillo, M., Peinado, A., and Beneyto, R. (2007). Determination of low analyte concentrations by near-infrared spectroscopy: effect of spectral pretreatments and estimation of multivariate detection limits. *Anal. Chim. Acta* 581, 318–323. doi: 10.1016/j.aca.2006.08.018

Chen, Q. S., Zhao, J. W., Liu, M. H., Cai, J. R., and Liu, J. H. (2008). Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms, *J. Pharma. Biomed.* 46, 568–573. doi: 10.1016/j.jpba.2007.10.031

Di, W., Yong, H., Nie, P. C., Fang, C., and Bao, Y. D. (2010). Hybrid variable selection in visible and near-infrared spectral analysis for non-invasive quality determination of grape juice. *Anal. Chim. Acta* 659, 229–237. doi: 10.1016/j.aca.2009.11.045

Dyrby, M., Engelsen, S. B., Nørgaard, L., Bruhn, M., and Lundsbergnielsen, L. (2002). Chemometric quantitation of the active substance (Containing C=N) in a pharmaceutical tablet using Near-Infrared (NIR) transmittance and NIR FT-Raman spectra. *Appl. Spectrosc.* 56, 579-585. doi: 10.1366/000370202195 5358

Esbensen, K. H., Geladi, P., and Larsen, A. (2014). The RPD myth.… *NIR news* 25, 24–28. doi: 10.1255/nirn.1462

Faber, N. K. (1999). Multivariate sensitivity for the interpretation of the effect of spectral pretreatment methods on near-infrared calibration model predictions. *Anal. Chem.* 71, 557–565. doi: 10.1021/ac980415r

Fernández-Cabanás, V. M., Garrido-Varo, A., Olmo, J. G., Pedro, E. D., and Dardenne, P. (2007). Optimisation of the spectral pre-treatments used for Iberian pig fat NIR calibrations. *Chemometri. Intell. Lab.* 87, 104–112. doi: 10.1016/j.chemolab.2006.10.005

Galvão, R. K. H., Araújo, M. C. U., Martins, M. D. N., José, G. E., Pontes, M. J. C., Silva, E. C., et al. (2006). An application of subagging for the improvement of prediction accuracy of multivariate calibration models. *Chemometri. Intell. Lab.* 81, 60–67. doi: 10.1016/j.chemolab.2005.09.005

Kachrimanis, K., Braun, D. E., and Griesser, U. J. (2007). Quantitative analysis of paracetamol polymorphs in powder mixtures by FT-Raman spectroscopy and PLS regression. *J. Pharma. Biomed.* 43, 407–412. doi: 10.1016/j.jpba.2006.07.032

Leardi, R., and Nørgaard, L. (2004). Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions. *J. Chemometr.* 18, 486–497. doi: 10.1002/cem.893

Lim, J., Kim, G., Mo, C., Kim, M. S., Chao, K., Qin, J., et al. (2016). Detection of melamine in milk powders using near-infrared hyperspectral imaging combined with regression coefficient of partial least square regression model. *Talanta* 151, 183–191. doi: 10.1016/j.talanta.2016.01.035

Mahanty, B., Yoon, S. U., and Kim, C. G. (2016). Spectroscopic quantitation of tetrazolium formazan in nano-toxicity assay with interval-based partial least squares regression and genetic algorithm. *Chemometri. Intell. Lab.* 154, 16–22. doi: 10.1016/j.chemolab.2016.03.012

Munck, L., Nielsen, J. P., Møller, B., Jacobsen, S., Søndergaard, I., Engelsen, S. B., et al. (2001). Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics. *Anal. Chim. Acta* 446, 169–184. doi: 10.1016/S0003-2670(01)01056-X

Pan, X. N., Li, Y., Wu, Z. S., Zhang, Q., Zheng, Z., Shi, X. Y., et al. (2015). A online NIR sensor for the pilot-scale extraction process in *Fructus aurantii* coupled with single and ensemble methods. *Sensors* 15, 8749–8763. doi: 10.3390/s150408749

Sarkhosh, M., Khorshidi, N., Niazi, A., and Leardi, R. (2014). Application of genetic algorithms for pixel selection in multivariate image analysis for a QSAR study of trypanocidal activity for quinone compounds and design new quinone compounds. *Chemometri. Intell. Lab.* 139, 168–174. doi: 10.1016/j.chemolab.2014.09.004

Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., Norgaard, L., and Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* 54, 413–419. doi: 10.1366/000370200194 9500

Üstün, B., Sanders, K. B., Dani, P., and Kellenbach, E. R. (2011). Quantification of chondroitin sulfate and dermatan sulfate in danaparoid sodium by 1H NMR spectroscopy and PLS regression. *Anal. Bioanal. Chem.* 399, 629–634. doi: 10.1007/s00216-010-4193-7

Williams, P.,(2014). Tutorial: the RPD statistic: a tutorial note. *NIR news* 25, 22–26. doi:10.1255/nirn.1419

Wu, Z. S., Du, M., Sui, C. L., Shi, X. Y., and Qiao, Y. J. (2012). Development and validation of nir model using low-concentration calibration range: rapid analysis of *Lonicera japonica* solution in ethanol precipitation process. *Anal. Methods* 4, 1084–1088. doi: 10.1039/C2AY05607K

Wu, Z. S., Ma, Q., Lin, Z. Z., Peng, Y. F., Ai, L., Shi, X. Y., et al. (2013a). A novel model selection strategy using total error concept. *Talanta* 107, 248–254. doi: 10.1016/j.talanta.2012.12.057

Wu, Z. S., Peng, Y. F., Chen, W., Xu, B., Ma, Q., Shi, X. Y., et al. (2013b). NIR spectroscopy as a process analytical technology (PAT) tool for monitoring and understanding of a hydrolysis process. *Bioresou. Technol.* 137, 394–399. doi: 10.1016/j.biortech.2013.03.008

Yu, F. L., Zhao, N., Wu, Z. S., Huang, M., Wang, D., Zhang, Y. B., et al. (2017). NIR rapid assessments of *Blumea balsamifera* (Ai-na-xiang) in China. *Molecules* 22:E1730. doi: 10.3390/molecules22101730

Zhao, N., Wu, Z. S., Zhang, Q., Shi, X. Y., Ma, Q., and Qiao, Y. J. (2015). Optimization of parameter selection for partial least squares model development. *Sci. Rep.* 5:11647. doi: 10.1038/srep11647

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 July 2018
doi: 10.3389/fchem.2018.00257

# Fusion of MALDI Spectrometric Imaging and Raman Spectroscopic Data for the Analysis of Biological Samples

*Oleg Ryabchykov*[1,2], *Juergen Popp*[1,2] *and Thomas Bocklitz*[1,2]*

[1] Spectroscopy and Imaging Research Department, Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, Jena, Germany, [2] Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Jena, Germany

Despite of a large number of imaging techniques for the characterization of biological samples, no universal one has been reported yet. In this work, a data fusion approach was investigated for combining Raman spectroscopic data with matrix-assisted laser desorption/ionization (MALDI) mass spectrometric data. It betters the image analysis of biological samples because Raman and MALDI information can be complementary to each other. While MALDI spectrometry yields detailed information regarding the lipid content, Raman spectroscopy provides valuable information about the overall chemical composition of the sample. The combination of Raman spectroscopic and MALDI spectrometric imaging data helps distinguishing different regions within the sample with a higher precision than would be possible by using either technique. We demonstrate that a data weighting step within the data fusion is necessary to reveal additional spectral features. The selected weighting approach was evaluated by examining the proportions of variance within the data explained by the first principal components of a principal component analysis (PCA) and visualizing the PCA results for each data type and combined data. In summary, the presented data fusion approach provides a concrete guideline on how to combine Raman spectroscopic and MALDI spectrometric imaging data for biological analysis.

**Keywords: MALDI-TOF, Raman imaging, data combination, data fusion, normalization, PCA**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Xia Guan,*
*Louisiana State University, United States*
*Ennio Carbone,*
*Università degli Studi Magna Græcia di Catanzaro, Italy*
*Frédéric Jacques Cuisinier,*
*Université de Montpellier, France*
*Anna V. Sharikova,*
*University at Albany, United States*

***Correspondence:**
*Thomas Bocklitz*
*thomas.bocklitz@uni-jena.de*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *17 December 2017*
***Accepted:*** *08 June 2018*
***Published:*** *16 July 2018*

***Citation:***
*Ryabchykov O, Popp J and Bocklitz T (2018) Fusion of MALDI Spectrometric Imaging and Raman Spectroscopic Data for the Analysis of Biological Samples. Front. Chem. 6:257. doi: 10.3389/fchem.2018.00257*

## INTRODUCTION

Different analytical methods could be utilized for biomedical analysis (e.g., cells, and tissues, etc.) to highlight a certain aspect of the sample e.g., morphological microstructure, distribution of electronic chromophores, molecule classes, or special proteins. Among the label-free imaging approaches, matrix-assisted laser desorption/ionization (MALDI) spectrometry, and Raman microscopy are certainly among the most powerful imaging techniques for the investigation of biomedical samples. Raman spectroscopy is a non-destructive spectroscopic method, which provides complex molecular information about the general chemical composition of the sample with a rather high spatial resolution (Abbe limit) to highlight subcellular features (Kong et al., 2015). The drawback of Raman imaging lies in its weak scattering efficiency that makes sampling time rather long for large area imaging. Raman spectroscopic imaging has

demonstrated its potential for biomedical diagnosis in numerous cancer-related studies (Tolstik et al., 2014), biological material analysis (Butler et al., 2016), cell characterization studies (Ramoji et al., 2012), and many other biomedical applications (Matousek and Stone, 2013; Ember et al., 2017).

On the other side, MALDI mass spectrometry provides information on specific substances, such as lipids or proteins (Fitzgerald et al., 1993). MALDI is a soft ionization technique utilized for mass-spectrometric imaging (Gessel et al., 2014) to determine large organic molecules and biomolecules undetected by conventional ionization techniques. This technique was employed in clinical parasitology (Singhal et al., 2016), microbial identification (Urwyler and Glaubitz, 2016), and cancer tissue investigation (Hinsch et al., 2017).

Raman spectroscopic and MALDI mass spectrometric imaging both offer a high molecular sensitivity. Moreover, Raman spectroscopy has been sequentially applied together with different mass spectrometric techniques to address a variety of biological tasks such as characterization of succinylated collagen (Kumar et al., 2011), investigation of microbial cells (Wagner, 2009), identification of fungal strains (Verwer et al., 2014) and characterization of lipid extracts from brain tissue (Köhler et al., 2009). In all the aforementioned studies, the Raman and mass spectrometric data are analyzed separately, and then summarized or compared to each other (Masyuko et al., 2014; Bocklitz et al., 2015; Muhamadali et al., 2016). To significantly increase the information content, Raman spectroscopic and MALDI mass spectrometric imaging data have to be co-registered (Bocklitz et al., 2013) followed by a high-level (distributed) data fusion. It means that each data type is analyzed separately to obtain the respective scores, which are then fused together. Alternatively, spectroscopic imaging can be used for mapping an area that is suitable for further investigation by means of MALDI spectrometric imaging (Fagerer et al., 2013) or a certain mass peak is used to define an area, from which the Raman spectra are analyzed (Bocklitz et al., 2013). Such a hierarchical pipeline corresponds to a decentralized data fusion approach.

In the present work, we introduced an analytical method to perform a low-level (centralized) fusion of Raman and MALDI imaging data. Because the experimental implementation of correlated imaging is challenging in many aspects (Masyuko et al., 2013), we utilized a computational approach to combine imaging data obtained by MALDI spectrometry and Raman spectroscopy. The correlation of Raman spectroscopy with mass spectrometric imaging techniques such as MALDI (Ahlf et al., 2014) or secondary ion mass spectrometry (SIMS) (Lanni et al., 2014) have proved its usefulness for biological applications. Moreover, a combination of MALDI imaging data with optical microscopy could attenuate instrumental effects (Van De Plas et al., 2015), and a joint analysis of vibrational and MALDI mass spectra could provide valuable information on brain tissue (Van De Plas et al., 2015; Lasch and Noda, 2017). Nevertheless, even if Raman and MALDI spectra are obtained by correlated imaging, each type of spectra shows its own specific features and should be preprocessed separately. Because the measurement techniques are based on different physical effects, the difference in data dimensionality and dynamic range can affect the contribution of each datatype in the analysis. Therefore, a weighting coefficient that balances the influence of Raman spectroscopic and MALDI spectrometric data in the data fusion center is required.

# MATERIALS AND METHODS

## Experimental Details

We demonstrated the data fusion on an example dataset of MALDI spectrometric and Raman spectroscopic scans obtained from the same mouse brain sample (*Mus musculus*) of 10 μm cryosection. The sample was cut on a cryostat, and then dried on a precooled conductive ITO-coated glass slide. Subsequently, Raman spectra were obtained using a confocal Raman microscope CRM-alpha300R (WITec, Ulm, Germany) and excited with a 633 nm HeNe laser (Melles Griot). The laser irradiation was adjusted in order to have about 10 mW power. The laser was coupled through an optical fiber into a Zeiss microscope. A spectral map was obtained by a raster scan with a 25 μm grid with a dwell time of 2 s and a pre-bleaching time of 1 s.

After the Raman scan, MALDI mass spectrometric imaging was performed with a common matrix alpha-cyano 4-hydroxy cinnamic acid (5 mg/mL) in 50% acetonitrile and 0.2% trifluoracetic acid. The ImagePrep station (Bruker Daltonics) was used to prepare and apply the matrix on the sample. The MALDI-time-of-flight (MALDI-TOF) spectrometric map was obtained on a Ultraflex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). A "smartbeam" laser ($\lambda = 355$ nm, repetition rate 200 Hz) was used. The spectrometer was calibrated with an external standard, a peptide calibration mixture (Bruker Daltonics). The measurements were performed in the positive reflectron mode with 500 shots per spectrum and spatial resolution of 75 μm.

Further experimental details for both data types and an example of a hierarchical data fusion implementation can be found in the report by Bocklitz et al. (2013). Nevertheless, in the context of a further discussion, it is important to highlight that in MALDI mass spectrometric imaging a matrix suitable for the analysis of the lipid content was applied.

## Preprocessing of Raman Spectroscopic Data

The influence of corrupting effects (e.g., cosmic spikes, fluorescence) on Raman spectra cannot be avoided completely. Thus, the development of complex preprocessing routines (Bocklitz et al., 2011) is required. To allow further analysis of the Raman spectra obtained with different calibrations, all spectra need to be interpolated to the same wavenumber axis (Dörfer et al., 2011). Moreover, keeping all the spectra in a single data matrix simplifies a further processing routine, so it is advantageous to perform the calibration as one of the first steps of the preprocessing workflow (**Figure 1**). Besides the wavenumber calibration, intensity calibration should be performed for the comparison of the measurements obtained with different devices

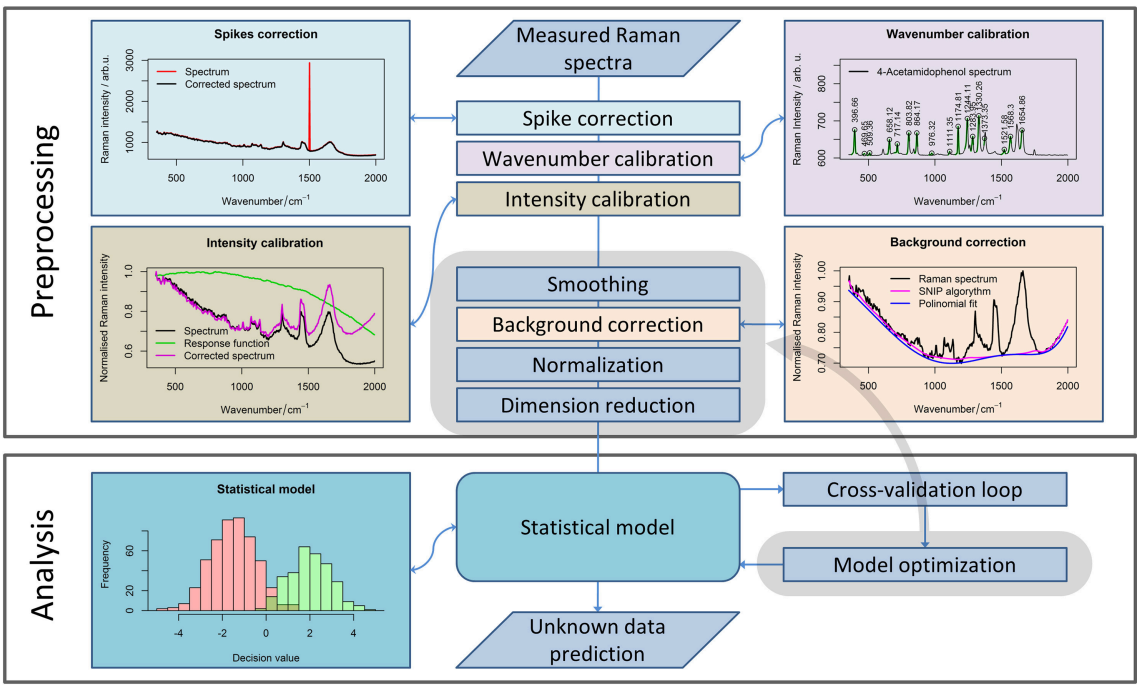



**FIGURE 1 |** A general pipeline for Raman data preprocessing. The workflow shows the main steps of the preprocessing routine necessary for robust Raman spectral analysis. Although some steps should be defined while planning the experiment, subsequent preprocessing methods (highlighted in gray area) and their parameters can be optimized for extracting the required information from the data.

or in the case where some changes in the measurement device have occurred (Dörfer et al., 2011).

The calibration is always needed for a reliable analysis, especially if the measurements were performed over a large time period, or settings of the device were changed between the measurements. In contrast, the following step within the preprocessing workflow (i.e., noise removal) is an optional step. However, among smoothing methods, only the running median with a relatively large window is applicable for cosmic ray noise removal. Unfortunately, filtering with a large window may corrupt the Raman bands themselves. Alternatively, 2–3 spectra per point can be acquired to eliminate the spikes that are not present in each spectrum. Nevertheless, this approach increases the measurement time dramatically. Therefore, this approach is not suitable for Raman imaging when a large number of spectra are recorded. Thus, specialized spike correction approaches like wavelet transform (Ehrentreich and Summchen, 2001), correlation methods (Cappel et al., 2010), calculation of the Laplacian of the spectral data matrix (Schulze and Turner, 2014; Ryabchykov et al., 2016), or a difference between the original and a smoothed spectrum (Zhang and Henson, 2007) must be used for spike removal.

The next step in the preprocessing workflow for Raman spectra is fluorescence background removal. In this work, the sensitive nonlinear iterative peak (SNIP) clipping algorithm (Ryan et al., 1988) was used for baseline estimation. The SNIP algorithm can be utilized for background estimation for a number of spectral measurements, like X-ray and mass spectra.

After baseline correction, the Raman spectra must be normalized (Afseth et al., 2006) to complete the basic preprocessing. There are several normalization approaches (e.g., vector normalization, normalization to integrated spectral intensity, or a single peak intensity value) that enhance the stability of the spectral data. In this work, we used vector normalization and $l_1$-normalization (Horn and Johnson, 1990) for Raman spectra. The difference between normalization to integrated spectral intensity and $l_1$-normalization is that the latter utilized absolute intensity values. As a result, the difference between both normalization approaches becomes more significant when negative values appear in the baseline corrected spectra due to noise or baseline correction artifacts.

## Preprocessing of MALDI Spectrometric Data

Although the measurement techniques themselves differ dramatically for Raman and MALDI mass spectroscopic imaging data, the preprocessing of these data has a lot in common. The m/z values are set according to an internal calibration and may "float" slightly from one measurement to another. Therefore, a phase correction along the m/z axis must be performed within the preprocessing workflow (**Figure 2**) to ensure that the spectra obtained in different measurements are comparable. For this purpose, it is advisable to use the stable intense peaks within the phase correction routine (Gu et al., 2006).

From a theoretical point of view, MALDI spectra should not feature a spectral background. Nevertheless, in measured MALDI

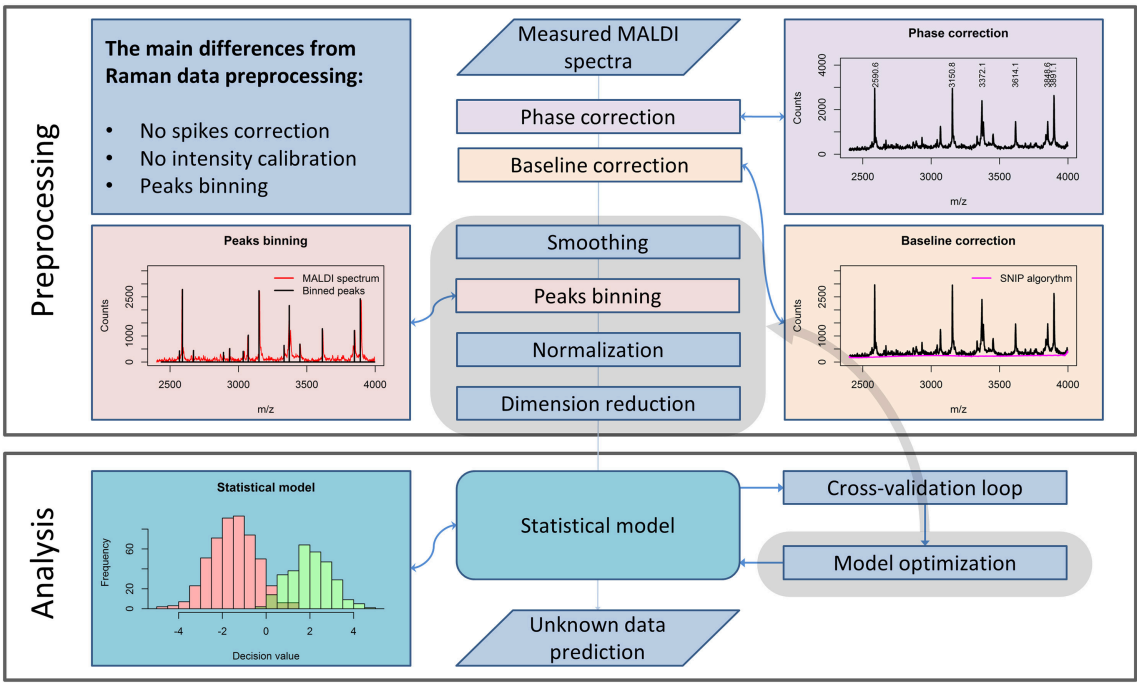



**FIGURE 2 |** A general pipeline for MALDI data preprocessing. The workflow shows the main steps of the preprocessing routine necessary for robust MALDI spectral data analysis and the main differences as compared to the Raman data preprocessing routine, described in **Figure 1**.

spectra a background is present. In literature, a background present in MALDI mass spectra is also known as "chemical noise background" (Krutchinsky and Chait, 2002). This type of noise results from matrix impurities and unstable ion clusters created during the sample scanning.

Similarly to Raman spectral preprocessing, the SNIP algorithm (Ryan et al., 1988) can be used to eliminate the background from mass spectra. Another complication in the analysis of MALDI spectra results from the fact that even after the phase correction, peak positions vary insignificantly among different spectra. An interpolation procedure, which is applied in Raman data preprocessing, would corrupt the sharp peaks found in MALDI spectra and is therefore not applied. To enable a direct comparison of the spectra, a binning procedure is applied. This procedure is based on the equalization of the m/z-values of peak positions within a certain range. Since the average peak width along the m/z axis increases with increased mass, the binning range is set with a so-called tolerance relative to the mass values. In contrast to Raman spectroscopy, intensity calibration for MALDI mass spectrometric imaging is not required. Nevertheless, normalization may be applied. Various types of normalization are used for MALDI mass spectroscopic imaging data: total ion count (TIC), vector norm (RMS), median, square root, logarithmic, and normalization to a noise level. In contrast to the Raman spectral data, MALDI mass spectra do not feature negative values. Thus, TIC normalization and normalization to $l_1$-norm, which is a sum of absolute values, are equal for MALDI spectra. If the significance level of the data is high, the normalization may be not necessary for the subsequent analysis.

## Computational Details

For MALDI data acquisition and calibration, a flexImaging software version 3.0 (Bruker Daltonics) was used. The data processing was also performed in R (R Core Team, 2017) using packages akima (Gebhardt)[1], Peaks (Morhac)[2], readBrukerFlexData (Gibb)[3], rsvd (Erichson)[4], spatstat (Baddeley and Turner, 2005), and Spikes (Ryabchykov et al., 2016).

Prior to the data preprocessing and data fusion, the MALDI and Raman spectra were interpolated to the same (spatial) grid by utilizing a co-registration framework. Based on the false color images of Raman spectroscopic and MALDI spectrometric scans, 6 points clearly representing the same positions on every scan were manually selected. The coordinates of the Raman spectroscopic map were then transformed to the coordinate system of the MALDI mass spectrometric map. Subsequently, the Raman spectra were interpolated to the grid of the MALDI mass spectral map. To perform this interpolation, every point within the Raman grid was assigned to the nearest point within the MALDI grid. After that, the average of the Raman spectra, assigned to the same point within the MALDI grid, was

[1]Gebhardt, H. A. "akima: Interpolation of Irregularly and Regularly Spaced Data."
[2]Morhac, M. "Peaks: Peaks."
[3]Gibb, S. "readBrukerFlexData: Reads Mass Spectrometry Data in Bruker *flex Format."
[4]Erichson, N. B. "rsvd: Randomized Singular Value Decomposition."

calculated. Two spectral maps were thus obtained and aligned in a point-wise manner.

After the alignment, the Raman spectroscopic and MALDI mass spectrometric imaging data were preprocessed. During the preprocessing, the wavenumber calibration of the Raman spectra and the phase correction of MALDI spectra were performed. The MALDI mass spectrometric imaging data were subsequently subjected to noise removal, background correction, and TIC normalization. The Raman spectra were corrected for fluorescence background and vector normalized. The SNIP algorithm was used for background estimation in both cases.

After the preprocessing, Raman and MALDI mass spectral data differed in their dimensionality and in dynamic range. Data with different dynamic ranges would contribute unequally in a further analysis and consequently the spectral matrices have to be additionally weighed before performing the PCA. The weighting coefficient was selected as a ratio between the $l_1$-norms of the matrices, which are sums over the absolute values in the matrix. After the weighting, the data were combined in a single matrix and analyzed with a PCA. To illustrate the benefit of data fusion and weighting, we also analyzed the un-weighted data in a combined manner and each data type separately. We also investigated the case, where the same normalization approach was applied to both data types and no additional weighting is required. When the Raman spectra were normalized to the total spectral intensity, which is equivalent to TIC normalization of mass spectra, the data matrices had equal $l_1$-norms.

## RESULTS AND DISCUSSION

Both Raman spectroscopic and MALDI mass spectrometric imaging data provide different insights into the chemical composition of the sample. Information on a broad range of molecules can be obtained from the Raman spectra. This information can be complemented by detailed information on lipid content, obtained from the MALDI data. To utilize both types of information together, a data fusion must be applied. This data fusion may be performed during different stages of the analysis workflow. Therefore, the architecture of the data processing workflow is dependent on the selected data fusion approach. These approaches can be divided into the following types (Castanedo, 2013):

- Centralized architecture (**Figure 3A**). The preprocessed data from different sources are combined in the data fusion center and are analyzed together.
- Decentralized architecture (**Figure 3B**). This scheme does not have a single data fusion center. The processing workflows are interacting at different processing stages. This architecture may provide multiple outputs or be represented as a hierarchical structure.
- Distributed architecture (**Figure 3C**). Each data type is preprocessed and analyzed separately. Subsequently, the output values are evaluated and combined to obtain a single result.

The decentralized and distributed architecture already showed their effectiveness for biomedical investigations (Bocklitz et al., 2013; Ahlf et al., 2014). The current work focuses on the centralized data fusion approach, also called low-level data fusion. In contrast to decentralized and distributed architectures, the centralized architecture shows a simpler workflow (**Figure 3A**). The data are combined in early steps of the analysis, directly after the preprocessing and even before the dimension reduction. At the data fusion center, where the different types of data are combined, an additional normalization or scaling of the data may be required to weight the influence of the different data types on the global model. The need for this weighting step arises from the differences in the data dimensionality, measurement units and dynamic ranges of the different measurement techniques. It is worth mentioning that the weighting is not a major issue in high-level data fusion approaches, which usually deal with standardized low-dimensional outputs of preliminary analysis in the data fusion center. However, a low-level data fusion (such as the applied centralized data fusion model) deals directly with preprocessed spectra of different types. Thus, the data scaling may dramatically influence extraction efficiency of the features.

To investigate the impact of data weighting, we searched for a marker that would allow an objective comparison of different data fusion and normalization approaches. This weighting scheme is designed for biological samples (i.e., a complex chemical composition), of which a large number of independent features have to be identified for appropriate description. By applying a PCA for dimension reduction, a large portion of the data variance is expected to be spread among multiple principal components (PCs) and the optimal approach should correspond to the slowest raise of the cumulative proportion of variance with a number of PCs.

The variances of the data explained by PCA are shown in the **Figure 4** where the normalization and fusion approaches (described in section Computational Details) are shown. Unfortunately, a direct comparison between cumulative proportions of variance obtained from Raman and MALDI mass spectral data, and their combined data is not suitable due to the different number of variables. However, different trends in the observed variance by the PCs in data with the same dimensionality can be interpreted. The left side of **Figure 4** shows that the variance of vector normalized Raman data is spread among a larger number of PCs than that of the total area normalized Raman data. This finding indicates that the vector normalization allows extracting a larger number of significant features from Raman data. Because the Raman spectra were vector normalized and the MALDI spectra were TIC normalized, the Raman data contribute more to the overall data variance than the MALDI data. Consequently, the PCA will focus on the variations in the Raman data and the variations in the MALDI data will have only a small influence. Alternatively, two datasets can be balanced by normalizing spectra of both types to their $l_1$-norms. By definition, this norm is a sum of absolute values. It takes dimensionality and scaling of the data into account, so no additional weighting is required. TIC normalization performed on MALDI data is already equal to $l_1$-normalization because

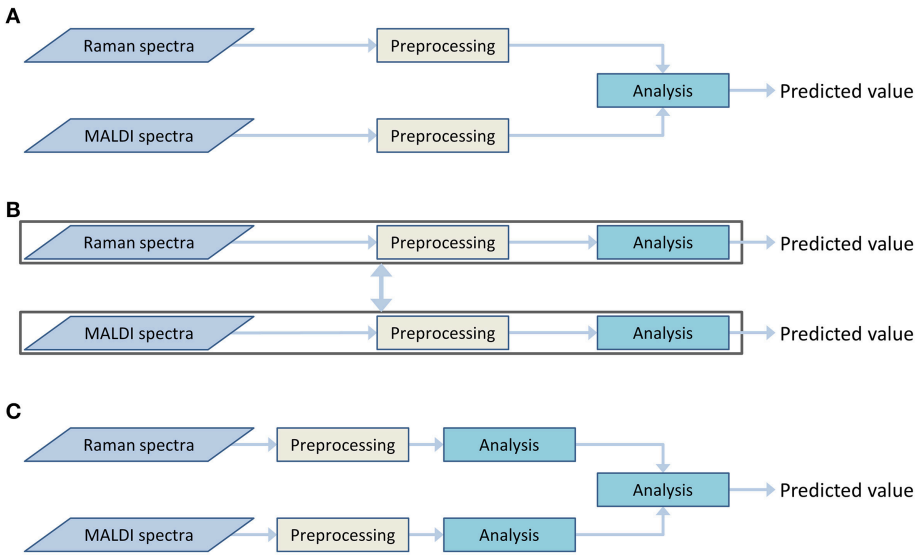



**FIGURE 3 |** Various data fusion architectures: centralized **(A)**, decentralized **(B)**, and distributed **(C)** architectures.

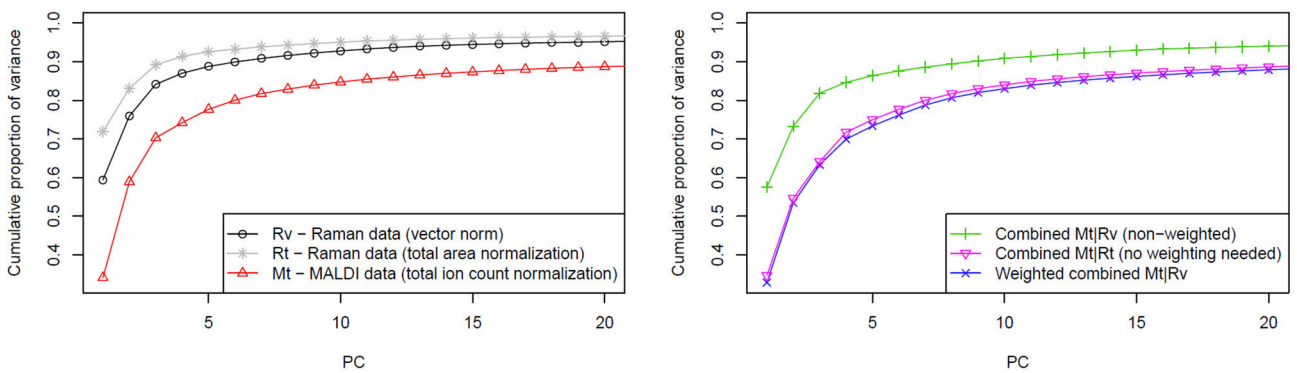



**FIGURE 4 |** Influence of weighting in the data fusion center on the PCA. The left side of the figure shows cumulative proportion of data variance explained by first 20 PCs for Raman data (normalized in two different ways) and for MALDI data. On the right side of the figure, a slower raise of cumulative proportion of weighted data variance in comparison to the non-weighted case is shown. This trend reflects that more independent features can be extracted from the data by applying weighting prior to the data fusion. As it is also shown in the plot on the right side, a similar effect can be reached by applying the same type of normalization for both data types.

there are no negative values present in the mass spectra. The right side of **Figure 4** clearly shows that there is a marked difference between the approach not taking the data scaling into account and the approaches based on weighting or identical normalization. However, no significant benefit was observed when comparing the weighting to identical normalization approach.

To further investigate the influence of weighting on data fusion, the weighting coefficient was varied in a range from 1 to 20 and a PCA utilized for every case. The extracted curves of the cumulative proportion of the variance were organized as a surface plot (**Figure 5**). To make the interpretation easier, the curves, which correspond to the data combination without weighting and with weighting based on the ratio of $l_1$-norms, are additionally highlighted in **Figure 5**. Although no single weighting coefficient is globally the best, the proposed weighting coefficient lies close to the area where the data variance is spread between multiple PCs. Thus, fusing data in this manner enables the PCA to extract a larger number of reliable features.

Although an optimal data fusion has been achieved as above-mentioned, a direct comparison of cumulative proportions of variance explained by the PCA for data with different dimensionalities may be misleading. Hence, the results obtained from the combined approach and separated data analysis (**Figure 6**) were checked by means of inspecting the PCA loadings and scores. The first three PCs were visualized separately for the MALDI spectrometric imaging data (**Figures 6A,C**), Raman spectroscopic imaging data (**Figures 6B,D**), and their combination (**Figures 6E–G**).

The comparison of the PCA scores in **Figure 6** shows that the image of the MALDI-Raman combination (**Figure 6G**) depicts clearer spatial features of the sample (compared to **Figures 6C,D**). The corresponding false-color score composite (**Figure 6G**) is less noisy, and looks subjectively better than the images obtained separately from the MALDI mass spectrometric (**Figure 6C**) and Raman spectroscopic data (**Figure 6D**). Moreover, the loading vector of the third PC of the MALDI spectra (shown in blue color in **Figure 6A**) has positive and negative values related to isotopes of the same molecules. It means that it represents mostly noise and variations in the signal to noise ratio. On the other hand, the MALDI part of the loadings of the third PC in the combined analysis (shown in blue color in **Figure 6E**) reflects a joint behavior for the isotopes of the same ions. Moreover, the Raman part of this PC contains the peaks associated with lipids (Notingher and Hench, 2006), namely the C = C stretching region (1,655–1,680 $cm^{-1}$), and CH deformation band (1,420–1,480 $cm^{-1}$). Although these two peaks may also be associated with Amide I and CH deformations of proteins, there is a decrease in the protein-associated range (Notingher and Hench, 2006) in the wavenumber region 1,128–1,284 $cm^{-1}$. Furthermore, there are notable changes in the CH-stretching region (2,800–3,100 $cm^{-1}$). Thus, the third PC of the combined data represents the actual diversity in the lipid composition of the sample. The relationship of the CH stretching region of the Raman spectra to the changes in the lipid content can also be observed by a high correlation of the Raman spectral region with MALDI mass spectra (**Figure 7**).

Since both data types simultaneously reflect variations in lipid content, the specific changes in the correlation profiles (**Figure 7**) of the Raman and MALDI data are observed in the areas related to lipid bands in Raman spectra. Besides the contributions of lipids, which are found in the third PC, the fingerprint region of Raman spectra contains numerous peaks related to proteins and DNA. These Raman bands correlate with MALDI peaks both positively and negatively (**Figure 7**). The correlation of a certain MALDI peak with the Raman data shows a similar structure, but with an opposite sign. This sign change reflects changes in the contribution of specific lipids with respect to the overall increase of lipid content in the sample.

One of the non-lipid compounds, which feature strong Raman bands, is phenylalanine. Its symmetrical ring breathing mode and C-H in-plane mode are visible in the first two PCs at 1,004 and 1,030 $cm^{-1}$. Another peak related to phenylalanine can be found in the first two PCs at 1,104 $cm^{-1}$ (Movasaghi et al., 2007). Aside of that, the first PC contains contributions of tryptophan at 760 $cm^{-1}$ (Bonifacio et al., 2010). The protein backbone C-$C_\alpha$ stretching of collagen is present in the second PC at 936 $cm^{-1}$ and the $\nu$(C–C) protein backbone is located in the first two PCs at 816 $cm^{-1}$ (Bonifacio et al., 2010). Also, prominent collagen-associated bands like Amide I and Amide III can be seen in the first PC at 1,655–1,680 and 1,220–1,284 $cm^{-1}$, respectively (Krafft et al., 2005; Notingher and Hench, 2006). Moreover, the peak at 1,647 $cm^{-1}$ is associated with the random coil structure of proteins in general (Movasaghi et al., 2007). This peak is also present in the first two PCs.

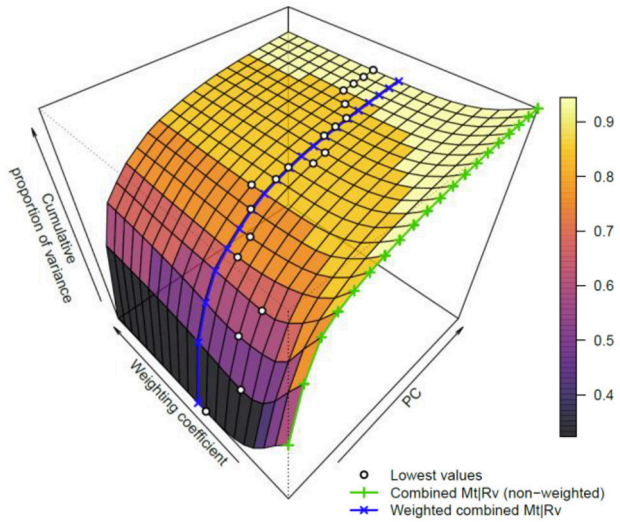



**FIGURE 5 |** Dependence of the variance explained by PCA using the weighting scheme. The surface plot covers the first 20 PCs and weighting coefficients between 1 and 20. The cumulative proportions of variance for the weighted and non-weighted cases are shown as blue and green lines, respectively (please refer to the online version for colors). Furthermore, the lowest variance is highlighted for each number of PCs with a dot. These dots represent an optimal unmixing for the related number of PCs. Although this optimum changes with respect to PC numbers, the used weighting coefficient based on $l_1$-norms clearly lies near the minimum of cumulative proportion of variance for a given number of PCs.

The main contribution to the first PC is the ratio between the fingerprint region of Raman spectra and C-H stretching region. On the other side, the fingerprint region of the second PC contains both positive and negative peaks, reflecting the changes in protein content. Along with the protein content, valuable information about DNA is obtained from the first two PCs of the Raman spectra. The peak at 1,180 $cm^{-1}$ represents cytosine and guanine. Another DNA peak is located at 1,263 $cm^{-1}$ and represents adenine and thymine (Movasaghi et al., 2007). All Raman spectral features provide a complex overview of the chemical composition of the mouse brain section. The MALDI data, on the other hand, extends the overview of the distribution of biomolecules based on Raman spectroscopy with detailed information about the lipid content composition.

## CONCLUSION

In this paper, a data fusion scheme was investigated to analyze Raman spectroscopic and MALDI mass spectrometric imaging data together. We described the most significant corrupting effects influencing the analysis of Raman spectroscopic and MALDI mass spectrometric imaging data. The preprocessing workflows were shown for the suppression of these corrupting effects by means of calibration, noise reduction, background correction, and normalization for both data types. After the pretreatment steps, the importance of data weighting prior to data fusion is highlighted, especially when the data are

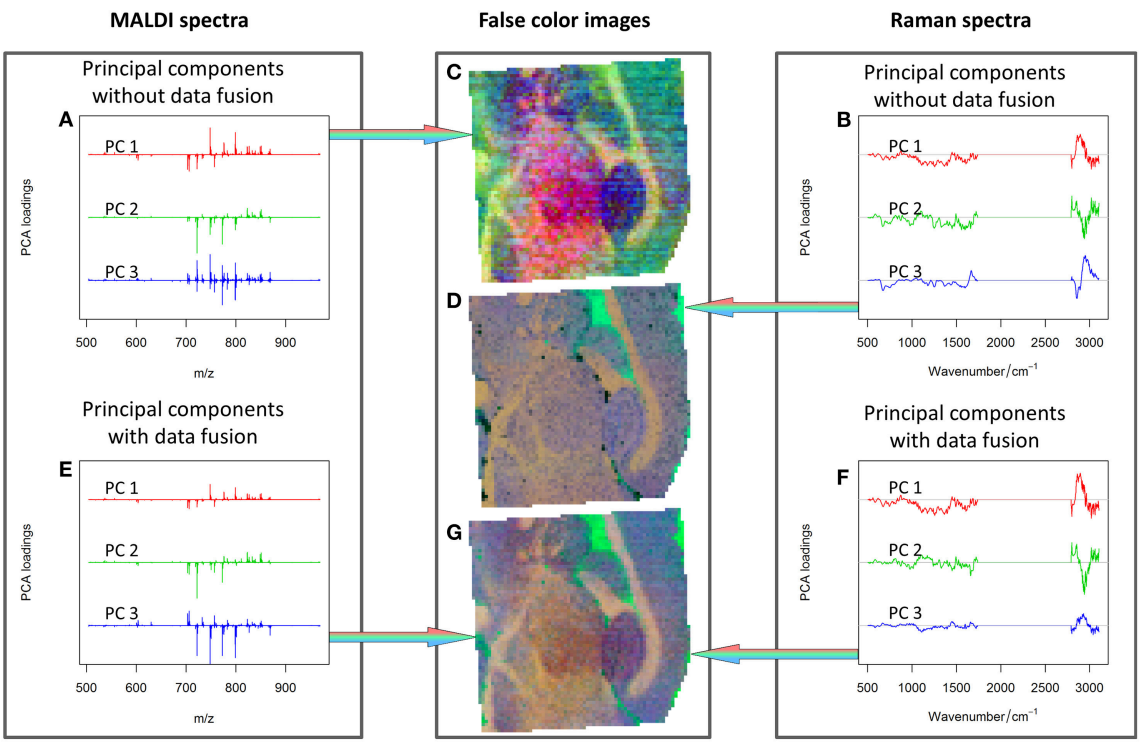



**FIGURE 6 |** PCA analysis: first three PCs calculated for MALDI spectra **(A)**, Raman spectra **(B)**, combined Raman-MALDI data **(E,F)** and their false-color score composites **(C,D,G)**. Red, green, and blue colors indicate the first, second and third PCs, respectively. Separate plots for the loadings and false color images can be found as Supplementary Material. The PCs composite image of the combined data **(G)** shows a smoother appearance, and the loadings after data fusion **(E,F)** are easier to interpret. See text for further details.

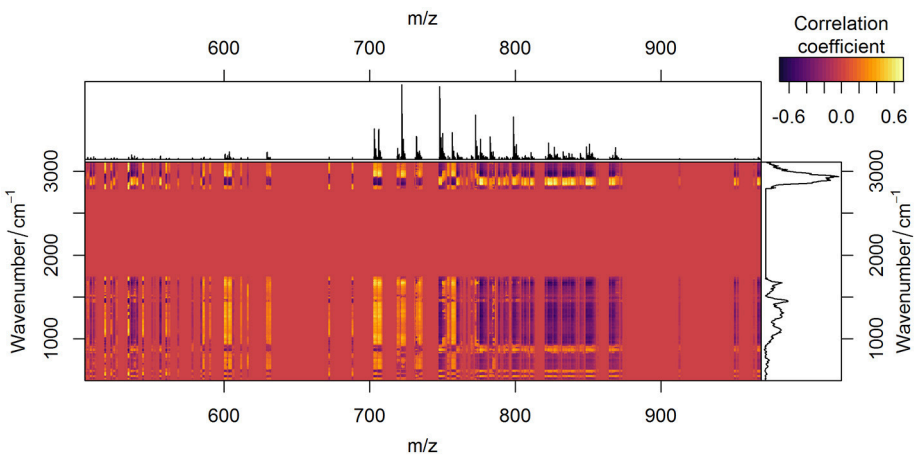



**FIGURE 7 |** Correlation between Raman spectroscopic and MALDI mass spectrometric data. Correlation of two data types after being preprocessed is depicted in yellow (positive values), red (zero), and violet (negative values) colors. Average preprocessed MALDI spectrum (on the top of the figure) and Raman spectrum (on the right side of the figure) are plotted for easier interpretation.

obtained from different sources and have different scales and dimensionalities. As there is no universal way of balancing the influence of data types on the analysis, optimization, and validation of weighting approaches should be done according to the specific data. In order to allow a judgment of the quality of a weighting, we proposed an approach that allows estimating the goodness of data weighting. This approach is based on analyzing proportions of data variance explained by PCs and we applied this approach by examining the cumulative variance. It was shown that the weighting, based on the ratio

of $l_1$-norms of the data matrices, allows optimal unmixing of the example data set into features. Besides the comparison of different weighting schemes, the proposed method can be used for the comparison of normalization approaches. It was found that vector normalization allows better unmixing of the example Raman data as compared to the normalization to the integrated spectral intensity ($l_1$-norm). Besides the establishment of a weighting approach, we discovered that a nearly optimal result compared to the weighting is achieved if the spectra of both types are normalized to the same norm. We could demonstrate this by normalizing both types of spectra of an example dataset to the same norm. This was the $l_1$-norm in our example. However, it is important to keep in mind that this method of comparing the cumulative proportions of variance should be used only when a researcher is interested in maximizing the number of extracted independent features.

The revealing of additional meaningful features by means of optimal data fusion was demonstrated for the combination of Raman spectroscopic and MALDI mass spectrometric imaging data. We showed this by comparing the third PC extracted from each type of data separately and from the combined data. The MALDI-related part of the third combined component showed a clearer interpretation in comparison to the third loading obtained from the MALDI data alone. Moreover, the Raman-related part of the combined component reflected variations in lipid to protein ratio. This PC depicts a decrease in a protein-associated range that occurs along with an increase of bands related to the CH deformation and C=C stretching in lipids, which can be found in the regions 1,128–1,284, 1,420–1,480, and 1,655–1,680 $cm^{-1}$, respectively. Therefore, changes in the lipid to protein ratio and changes in lipid content itself can be observed simultaneously through the data fusion of Raman spectroscopic and MALDI mass spectrometric imaging data.

Finally, the advantage of the combined analysis was illustrated by a comparison of the PCA results visualized as false-color RGB images. These images were obtained separately for the preprocessed Raman and MALDI imaging data and for the combined data. Visual investigation of the images showed that the combined approach provides a sharper image with less noise contributions. This allows the conclusion that the data fusion increases reliability not only for the spectral but also for the spatial features present in the data.

## ETHICS STATEMENT

This research is based on already published data provided to the authors by Bocklitz et al. (2013). For this reason, an ethics approval was not required as per institutional and national guidelines.

## AUTHOR CONTRIBUTIONS

TB and JP initiated the study, supervised the study and discussed the results. OR performed the analysis including the development of the R scripts. TB performed the pre-study including the co-registration step. OR, JP, and TB wrote the manuscript.

## ACKNOWLEDGMENTS

Financial support of the EU via the project HemoSpec (FP 7, CN 611682), co-funding of the EU for the project PhotoSkin (FKZ 13N13243) and support of the BMBF via the project PhotoSkin (FKZ 13N13243) and Uro-MDD (FKZ 03ZZ0444J) are highly acknowledged. The publication of this article was funded by the Open Access Fund of the Leibniz Association. Authors wish to thank Dr. Anna Crecelius for acquiring the data and to Prof. Dr. Ferdinand von Eggeling for helpful discussions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00257/full#supplementary-material

**Supplementary Image 1 |** The plots from Figure 6 provided in vector format.

## REFERENCES

Afseth, N. K., Segtnan, V. H., and Wold, J. P. (2006). Raman spectra of biological samples: a study of preprocessing methods. *Appl. Spectrosc.* 60, 1358–1367. doi: 10.1366/000370206779321454

Ahlf, D. R., Masyuko, R. N., Hummon, A. B., and Bohn, P. W. (2014). Correlated mass spectrometry imaging and confocal Raman microscopy for studies of three-dimensional cell culture sections. *Analyst* 139, 4578–4585. doi: 10.1039/C4AN00826J

Baddeley, A., and Turner, R. (2005). spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* 12:42. doi: 10.18637/jss.v012.i06

Bocklitz, T., Bräutigam, K., Urbanek, A., Hoffmann, F., Von Eggeling, F., Ernst, G., et al. (2015). Novel workflow for combining Raman spectroscopy and MALDI-MSI for tissue based studies. *Anal. Bioanal. Chem.* 407, 7865–7873. doi: 10.1007/s00216-015-8987-5

Bocklitz, T., Walter, A., Hartmann, K., Rösch, P., and Popp, J. (2011). How to pre-process Raman spectra for reliable and stable models? *Anal. Chim. Acta* 704, 47–56. doi: 10.1016/j.aca.2011.06.043

Bocklitz, T. W., Crecelius, A. C., Matthäus, C., Tarcea, N., Von Eggeling, F., Schmitt, M., et al. (2013). Deeper understanding of biological tissue: quantitative correlation of MALDI-TOF and Raman imaging. *Anal. Chem.* 85, 10829–10834. doi: 10.1021/ac402175c

Bonifacio, A., Beleites, C., Vittur, F., Marsich, E., Semeraro, S., Paoletti, S., et al. (2010). Chemical imaging of articular cartilage sections with Raman mapping, employing uni- and multi-variate methods for data analysis. *Analyst* 135, 3193–3204. doi: 10.1039/c0an00459f

Butler, H. J., Ashton, L., Bird, B., Cinque, G., Curtis, K., Dorney, J., et al. (2016). Using Raman spectroscopy to characterize biological materials. *Nat. Protocols* 11, 664–687. doi: 10.1038/nprot.2016.036

Cappel, U. B., Bell, I. M., and Pickard, L. K. (2010). Removing cosmic ray features from Raman map data by a refined nearest neighbor comparison method as a precursor for chemometric analysis. *Appl. Spectro.* 64, 195–200. doi: 10.1366/000370210790619528

Castanedo, F. (2013). A review of data fusion techniques. *Sci. World J.* 2013:19. doi: 10.1155/2013/704504

Dörfer, T., Bocklitz, T., Tarcea, N., Schmitt, M., and Popp, J. (2011). Checking and improving calibration of Raman spectra using chemometric approaches. *Zeitschrift Fur Phys. Chem.* 225, 753–764. doi: 10.1524/zpch.2011.0077

Ehrentreich, F., and Sümmchen, L. (2001). Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal. Chem.* 73, 4364–4373. doi: 10.1021/ac0013756

Ember, K. J. I., Hoeve, M. A., McAughtrie, S. L., Bergholt, M. S., Dwyer, B. J., Stevens, M. M., et al. (2017). Raman spectroscopy and regenerative medicine: a review. *Regenerat. Med.* 2:12. doi: 10.1038/s41536-017-0014-3

Fagerer, S. R., Schmid, T., Ibáñez, A. J., Pabst, M., Steinhoff, R., Jefimovs, K., et al. (2013). Analysis of single algal cells by combining mass spectrometry with Raman and fluorescence mapping. *Analyst* 138, 6732–6736. doi: 10.1039/c3an01135f

Fitzgerald, M. C., Parr, G. R., and Smith, L. M. (1993). Basic matrixes for the matrix-assisted laser desorption/ionization mass spectrometry of proteins and oligonucleotides. *Anal. Chem.* 65, 3204–3211. doi: 10.1021/ac00070a007

Gessel, M. M., Norris, J. L., and Caprioli, R. M. (2014). MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *J. Prot.* 107, 71–82. doi: 10.1016/j.jprot.2014.03.021

Gu, M., Wang, Y., Zhao, X. G., and Gu, Z. M. (2006). Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system. *Rapid Commun. Mass Spectrosc.* 20, 764–770. doi: 10.1002/rcm.2377

Hinsch, A., Buchholz, M., Odinga, S., Borkowski, C., Koop, C., Izbicki, J. R., et al. (2017). MALDI imaging mass spectrometry reveals multiple clinically relevant masses in colorectal cancer using large-scale tissue microarrays. *J. Mass Spectro.* 52, 165–173. doi: 10.1002/jms.3916

Horn, R. A., and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press.

Köhler, M., Machill, S., Salzer, R., and Krafft, C. (2009). Characterization of lipid extracts from brain tissue and tumors using Raman spectroscopy and mass spectrometry. *Anal. Bioanal. Chem.* 393, 1513–1520. doi: 10.1007/s00216-008-2592-9

Kong, K., Kendall, C., Stone, N., and Notingher, I. (2015). Raman spectroscopy for medical diagnostics — From *in-vitro* biofluid assays to *in-vivo* cancer detection. *Adv. Drug Deliv. Rev.* 89, 121–134. doi: 10.1016/j.addr.2015.03.009

Krafft, C., Knetschke, T., Funk, R. H. W., and Salzer, R. (2005). Identification of organelles and vesicles in single cells by Raman microspectroscopic mapping. *Vibrat. Spectrosc.* 38, 85–93. doi: 10.1016/j.vibspec.2005.02.008

Krutchinsky, A. N., and Chait, B. T. (2002). On the nature of the chemical noise in MALDI mass spectra. *J. Am. Soc. Mass Spectrosc.* 13, 129–134. doi: 10.1016/S1044-0305(01)00336-1

Kumar, R., Sripriya, R., Balaji, S., Senthil Kumar, M., and Sehgal, P. K. (2011). Physical characterization of succinylated type I collagen by Raman spectra and MALDI-TOF/MS and *in vitro* evaluation for biomedical applications. *J. Mol. Struct.* 994, 117–124. doi: 10.1016/j.molstruc.2011.03.005

Lanni, E. J., Masyuko, R. N., Driscoll, C. M., Dunham, S. J. B., Shrout, J. D., Bohn, P. W., et al. (2014). Correlated imaging with C60-SIMS and confocal raman microscopy: visualization of cell-scale molecular distributions in bacterial biofilms. *Anal. Chem.* 86, 10885–10891. doi: 10.1021/ac5030914

Lasch, P., and Noda, I. (2017). Two-dimensional correlation spectroscopy for multimodal analysis of FT-IR, Raman, and MALDI-TOF MS hyperspectral images with Hamster brain tissue. *Anal. Chem.* 89, 5008–5016. doi: 10.1021/acs.analchem.7b00332

Masyuko, R., Lanni, E. J., Sweedler, J. V., and Bohn, P. W. (2013). Correlated imaging - a grand challenge in chemical analysis. *Analyst* 138, 1924–1939. doi: 10.1039/c3an36416j

Masyuko, R. N., Lanni, E. J., Driscoll, C. M., Shrout, J. D., Sweedler, J. V., and Bohn, P. W. (2014). Spatial organization of *Pseudomonas aeruginosa* biofilms probed by combined matrix-assisted laser desorption ionization mass spectrometry and confocal Raman microscopy. *Analyst* 139, 5700–5708. doi: 10.1039/C4AN00435C

Matousek, P., and Stone, N. (2013). Recent advances in the development of Raman spectroscopy for deep non-invasive medical diagnosis. *J. Biophot.* 6, 7–19. doi: 10.1002/jbio.201200141

Movasaghi, Z., Rehman, S., and Rehman, I. U. (2007). Raman spectroscopy of biological tissues. *Appl. Spectro. Rev.* 42, 493–541. doi: 10.1080/05704920701551530

Muhamadali, H., Weaver, D., Subaihi, A., Almasoud, N., Trivedi, D. K., Ellis, D. I., et al. (2016). Chicken, beams, and Campylobacter: rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry. *Analyst* 141, 111–122. doi: 10.1039/C5AN01945A

Notingher, I., and Hench, L. L. (2006). Raman microspectroscopy: a noninvasive tool for studies of individual living cells *in vitro*. *Exp. Rev. Med. Dev.* 3, 215–234. doi: 10.1586/17434440.3.2.215

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramoji, A., Neugebauer, U., Bocklitz, T., Foerster, M., Kiehntopf, M., Bauer, M., et al. (2012). Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood. *Anal. Chem.* 84, 5335–5342. doi: 10.1021/ac3007363

Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Foerster, M., Kroegel, C., et al. (2016). Automatization of spike correction in Raman spectra of biological samples. *Chemometr. Intell. Lab. Syst.* 155, 1–6. doi: 10.1016/j.chemolab.2016.03.024

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., and Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instr. Methods Phys. Res. B* 34, 396–402. doi: 10.1016/0168-583X(88)90063-8

Schulze, H. G., and Turner, R. F. (2014). A two-dimensionally coincident second difference cosmic ray spike removal method for the fully automated processing of Raman spectra. *Appl. Spectro.* 68, 185–191. doi: 10.1366/13-07216

Singhal, N., Kumar, M., and Virdi, J. S. (2016). MALDI-TOF MS in clinical parasitology: applications, constraints and prospects. *Parasitology* 143, 1491–1500. doi: 10.1017/S0031182016001189

Tolstik, T., Marquardt, C., Matthäus, C., Bergner, N., Bielecki, C., Krafft, C., et al. (2014). Discrimination and classification of liver cancer cells and proliferation states by Raman spectroscopic imaging. *Analyst* 139, 6036–6043. doi: 10.1039/C4AN00211C

Urwyler, S. K., and Glaubitz, J. (2016). Advantage of MALDI-TOF-MS over biochemical-based phenotyping for microbial identification illustrated on industrial applications. *Lett. Appl. Microbiol.* 62, 130–137. doi: 10.1111/lam.12526

Van De Plas, R., Yang, J., Spraggins, J., and Caprioli, R. M. (2015). Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nat. Methods* 12:366. doi: 10.1038/nmeth.3296

Verwer, P. E., Van Leeuwen, W. B., Girard, V., Monnin, V., Van Belkum, A., Staab, J. F., et al. (2014). Discrimination of Aspergillus lentulus from *Aspergillus fumigatus* by Raman spectroscopy and MALDI-TOF MS. *Eur. J. Clin. Microbiol. Infect. Dis.* 33, 245–251. doi: 10.1007/s10096-013-1951-4

Wagner, M. (2009). Single-cell ecophysiology of microbes as revealed by Raman microspectroscopy or secondary ion mass spectrometry imaging. *Ann. Rev. Microbiol.* 63, 411–429. doi: 10.1146/annurev.micro.091208.073233

Zhang, L., and Henson, M. J. (2007). A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Appl. Spectro.* 61, 1015–1020. doi: 10.1366/000370207781745847

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 31 July 2018
doi: 10.3389/fchem.2018.00325

# Real-Time Analysis of Potassium in Infant Formula Powder by Data-Driven Laser-Induced Breakdown Spectroscopy

*Da Chen, Jing Zong, Zhixuan Huang, Junxin Liu and Qifeng Li**

*College of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin, China*

Potassium represents one of the most crucial minerals in infant formula that supports healthy growth and development of infants. Here, a novel strategy for the real-time quantification of potassium in infant formula samples is introduced. Using laser-induced breakdown spectroscopy (LIBS) in a data-driven approach, a modified random frog algorithm (MRFA) is adopted in a higher-density discrete wavelet transform (HDWT) domain for the selection of the most important features related to potassium, which is named as DD-LIBS. In DD-LIBS, the HDWT oversamples the LIBS signals in both time and frequency domains by a factor of two, enhancing the spectral expandability in an approximately shift-invariant way. The MRFA is thus capable of isolating the features of potassium with experience accumulated from the collected LIBS data. Such pretreatment combined with a partial least squared (PLS) model can significantly suppress the uncontrolled shift and broadening effects on multivariate calibration, improving the capability of LIBS for accurate quantification of potassium. The present work demonstrates the feasibility of DD-LIBS for the quantification of potassium content of 90 commercial infant formula samples. A satisfactory result illustrates DD-LIBS as a feasible tool for real-time analysis of potassium content with little sample preparation. This strategy may be well extended to other element detection in the presence of uncontrolled interference.

**Keywords: laser-induced breakdown spectroscopy, higher density wavelet transform, modified random frog algorithm, infant formula, potassium**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Venugopal Rao Soma,*
*University of Hyderabad, India*
*Noureddine Melikechi,*
*University of Massachusetts Lowell, United States*

****Correspondence:***
*Qifeng Li*
*qfli@tju.edu.cn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *28 February 2018*
***Accepted:*** *11 July 2018*
***Published:*** *31 July 2018*

***Citation:***
*Chen D, Zong J, Huang Z, Liu J and Li Q (2018) Real-Time Analysis of Potassium in Infant Formula Powder by Data-Driven Laser-Induced Breakdown Spectroscopy. Front. Chem. 6:325. doi: 10.3389/fchem.2018.00325*

## INTRODUCTION

Infant formula, as a breast-milk substitute, plays a significant role since it is the sole source of nutrition for some infants (Deckelbaum et al., 2004; Meucci et al., 2010; Codex, 2015; AOAC International, 2016). The international standard for infant formula set by Codex Alimentarius Commission (CAC) has a strict requirement of the essential composition and nutrition content (Codex, 2015). Meanwhile, all infant formulas marketed must also meet local standards, which are based on the national physique and health level (The Ministry of Health People's Republic of China, 2010b). As an essential cation in intracellular fluid, potassium is one

**Abbreviations:** LIBS, Laser-induced breakdown spectroscopy; RFA, random frog algorithm; MRFA, modified random frog algorithm; HDWT, higher density wavelet transform; PLS, partial least square.

of the most important minerals to support healthy growth and development of infants, because potassium is critically involved with acid-based balance function, osmotic pressure regulation, nerve impulse conduction, muscle construction and $Na^+/K^+$ ATPase (Soetan et al., 2010). An incorrect intake of potassium can also cause diseases (such as hyperkalemia and hypokalemia), which therefore turns the correct control of potassium content of infant formula into a superior importance for both international and local standards (Deckelbaum et al., 2004; Koletzko et al., 2005; The Ministry of Health People's Republic of China, 2010b; Codex, 2015).

To determine the potassium content, the current standard analytical methods are mostly based on atomic absorption spectrophotometry (AAS) (The Ministry of Health People's Republic of China, 2010a), inductively coupled plasma atomic emission spectrometry (ICP-AES) (The Ministry of Health People's Republic of China, 2010a; ISO, 2018a) and inductively coupled plasma mass spectrometry (ICP-MS) (ISO, 2018b), etc. These methods require a laborious and time-consuming sample processing procedure, together with strictly controlled laboratory environment and large sample volume (Panne et al., 2001; Awan et al., 2013; Matsumoto et al., 2016). However, the huge consumption of infant formula at a level of million tons greatly challenges the efficiency of current analytical methods (Tan et al., 2017), and leads to the necessity to develop an efficient and simple method for quantifying the potassium content in infant formula.

Laser-induced breakdown spectroscopy (LIBS), an optical emission spectroscopy technique, presents a potential solution to this challenge (Aragón and Aguilera, 2008). In LIBS, a high-power density laser pulse is focused on a target material in less than a nanosecond, during which a high-temperature plasma is generated by vaporizing a small portion of the target (Zheng et al., 2014). As a result, the radiant characteristics of elements are emitted by the excited atomic, ionic, and molecular fragments produced by the plasma (Harmon et al., 2006; Bousquet et al., 2007). Hence, LIBS offers a strong capability to rapidly detect the element contents in many type of samples (Panne et al., 2001; Bousquet et al., 2007; Hussain and Gondal, 2008; Eseller et al., 2010), with little sample preparation (Hahn and Omenetto, 2010; Hou et al., 2016).

The development of lasers, optics and charge-coupled array detectors has driven a critical revolution in the sensitivity of LIBS, making it a "future superstar" analytical method (Hou et al., 2016). However, the complex process of laser-sample and plasma-particle interactions may distort LIBS peaks (Hahn and Omenetto, 2012). The spectral interference presented in the LIBS signals often leads to an unresolved, broadened and often shifted center of gravity that introduces wavelength shift of spectral peaks (Cremers and Radziemski, 2013), which compromises the LIBS calibration performance. Alternatively, a calibration-free LIBS (CF-LIBS) based on strict theoretical assumptions of laser induced plasma may estimate analyte concentrations correctly. However, CF-LIBS data are severely affected by the self-absorption effect and estimation of plasma temperature (Sun and Yu, 2009), which is challenging for pharmaceutical applications. To improve calibration results, the higher-density discrete wavelet (HDWT) signal processing method with shift-invariant capability becomes a good candidate (Selesnick, 2006). With HDWT, a minor wavelength shift in the raw spectra will not cause a significant variance of the HDWT coefficients at different scales (Qin et al., 2010), which guarantees the reliability of the future calibration models with the HDWT coefficients.

The unique feature of HDWT is that it processes the spectral data in an approximately shift-invariant way, while oversampling the spectral signals in both time and frequency domains by a factor of two, as opposed to the shift-variant downsampling in the conventional discrete wavelet transform (DWT) (Selesnick, 2006). It allows to generate triple wavelet coefficients and thus enables to isolate the localized LIBS spectral features more accurately and robustly (Han et al., 2017). After being processed by HDWT, the LIBS spectral bands of potassium can be well extracted by specific HDWT coefficients, which can be optimized by the feature selection methods (Yun et al., 2013). Since the underlying mechanism of LIBS signals is too complex to be interpreted directly, the observed LIBS data themselves must drive variable selection to optimize multivariate calibration (Parab et al., 2009).

Several feature selection procedures have been developed, including random frog algorithm (RFA) (Li et al., 2012), competitive adaptive reweighted sampling (CARS) (Li et al., 2009), uninformative variable elimination (UVE) and its derivation (Cai et al., 2008; Moros et al., 2008), and randomization tests (Kennedy and Cade, 1996) etc. Among above-mentioned procedures, RFA presents a unique advantage in processing high dimensional spectral data without any prior knowledge that matches the demand of data-driven well. However, the RFA tends to generate a semi-random result that may not correlate accurately with targeted chemicals. In this case, a modified random frog algorithm (MRFA) is adopted by the multiple resampling strategy, in which the RFA has executed hundreds of times to select variables with the highest probability. Therefore, the MRFA is expected to improve the reliability of the LIBS models.

In this work, a data-driven strategy is proposed to isolate the spectral features of potassium with experience accumulated from the observed LIBS data. This strategy aims to estimate the relationship between LIBS spectral datasets and potassium concentrations from the existing input-output data (Gani et al., 2009), which is named as data-driven LIBS (DD-LIBS). In DD-LIBS, the MRFA was adopted in the HDWT domains instead of raw LIBS spectra to avoid spectral interference. A calibration model was then constructed with the selected HDWT coefficients. The DD-LIBS strategy was validated by using 90 commercial infant formula samples.

# MATERIALS AND METHODS

## Sample Resource and Preparation

Samples of 90 commercially available infant formulas were purchased from the local market, which includes 24 mainstream brands in China. The potassium content was measured by flame atomic absorption spectrometry according to the Chinese national test standard method GB5009.91-2017. To reduce the effects of particle size on LIBS signals, solid infant

formula samples were pressed into compact pellets by using a hydraulic press machine under 30 MPa pressure. The measurable characteristics of diameter, thickness, and mass of the pellets were 20 mm, 10 mm, and 4 g, respectively.

## Laser-Induced Breakdown Spectrometry System

In this study, an Ocean Optics LIBS 2500-7 spectrometer system was equipped with CFR Nd. YAG Laser source (LIBS-LAS200MJ, Big Sky Laser Technologies). The laser was operated at a fundamental wavelength of 1,064 nm, and the pulse energy utilized in this experiment was 50 mJ. The pulse duration was 9.5 ns, and the pulse repetition rate was 10 Hz. The LIBS 2500-7 has seven channels to provide a broad spectral wavelength range from 200 to 880 nm, covering the emission spectra of all elements. Each channel is equipped with a 2048-element linear CCD array to present a high optical resolution of 0.1 nm (FWHM). The frame rate was 10 Hz. The integration time was 2.1 ms, and it could be changed in a free-run mode to match sample properties. The trigger delay was from −121 to +135 μs in 500 ns steps. The delay time was set at 0.83 μs, which was determined through optimizing the signal-background ratio (SBR) and characteristic spectral intensity.

## Experimental Procedure

For each LIBS analysis, the pellets were put on the sample stage, and 10 different spots of one pellet were evenly selected for LIBS measurement, which reduces the effects of inhomogeneity and surface variations on LIBS signals. Each spot was ablated with 10 laser pulses. As a result, total 100 LIBS spectra were collected and averaged into a single LIBS spectrum, which improves the stability of LIBS experiments.

## Calibration Approach

Samples were randomly divided into two sets, i.e., a 65-sample set was used to build a calibration model and a 25-sample set was used to validate the calibration model.

### Normalization Methods

In order to use LIBS in a timely manner, minimal sample pretreatment is preferred. Thus, in LIBS measurement, normalization is performed to compensate for physical variations and sample matrix differences. In this work, five normalization methods, such as average, normalization by norm, spectral area, spectral height, and carbon emission lines (Abdel-Salam et al., 2013; Castro and Pereirafilho, 2016; dos Santos Augusto et al., 2017), were compared.

### Data Analysis Through Data-Driven LIBS

The LIBS spectra are affected by matrix effect and other unknown interference, resulting in broadened and shifted LIBS peaks. DD-LIBS is thus proposed to reduce the effect of peak broadening and shift on multivariate calibration. To correct shifted and expanded spectral peaks, HDWT was applied by implementing the three channel filter banks to conduct an oversampling operation for generating nearly shift-invariant wavelet coefficients.

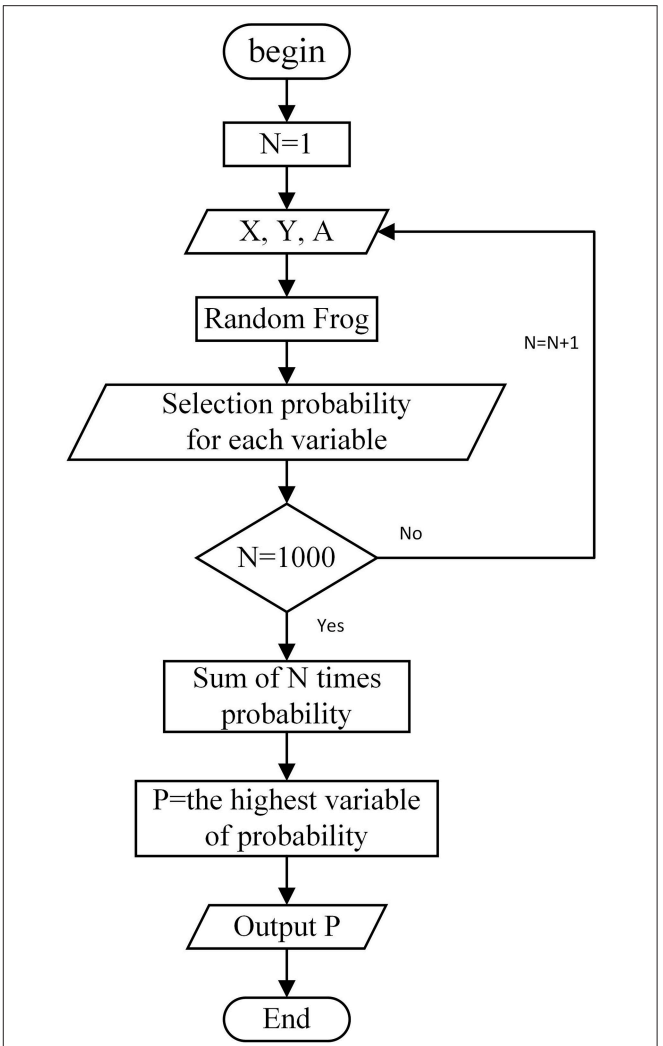



**FIGURE 1 |** Flowchart of modified random frog algorithm (MRFA), where X is the HDWT coefficients of calibration set, Y is the reference values of calibration samples, A is the number of PLS factors, N is the number of MRFA runs, P is the sum of the selected probability of each variable.

After the HDWT calculation, the raw LIBS spectra were decomposed into localized components labeled by a scale, facilitating the feature selection methods to isolate the spectral bands related to potassium. Then, the MRFA was performed by using the bagging strategy, assigning 70% samples to a training subset and 30% samples to a validation set. The procedure was repeated for 1,000 times to generate 1,000 different selection probabilities of each HDWT coefficient for accumulation. The flowchart of MRFA is shown in **Figure 1**.

In this work, only the HDWT coefficient with the highest probability was selected for further calibration because it provided valuable robustness against the uncontrolled and unknown spectral interference, and the feature selection result can be easily validated by the reference LIBS spectra of potassium.

As mentioned above, DD-LIBS was established by integrating HDWT, MRFA and PLS together. The HDWT codes were written

in Matlab 2013a based on the Selesnick's theory (Selesnick, 2006). The programs of PLS and RFA were available in the libPLS toolbox for Matlab (Li et al., 2014), and the MRFA was modified from RFA in Matlab 2013a.

### Evaluation Parameters

The root mean square error of cross-validation (*RMSECV*) was used to determine the HDWT parameters, and the coefficient of determination ($R^2$) was used to evaluate the calibration performance of the developed models (Chu, 2011):

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{m} (y_{i,actual} - y_{i,predicted})^2}{m-1}} \quad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{i,actual} - y_{i,predicted})^2}{\sum_{i=1}^{n} (y_{i,actual} - \bar{y}_{i,actual})^2} \quad (2)$$

Where $y_{i,actual}$ is the reference value of the potassium concentration of sample $i$, $y_{i,predicted}$ represents the predicted value of sample $i$, $m$ is the number of calibration samples, and $\bar{y}_{i,actual}$ represents the average reference concentration of all samples. When we obtain a *RMSECV* from the prediction set, we refer it as a *RMSEP*. The evaluation criterion is very simple: the smaller the value of *RMSEP* is, the stronger the prediction capability of the model is.

The limit of detection (LOD) was calculated by using the following equation (ICH Guideline, 2005):

$$\mathrm{LOD} = \frac{3.3 \times SD_{blank}}{s} \quad (3)$$

Where $SD_{blank}$ is the standard deviation of the baseline near peaks, and $s$ is the slope of the calibration curve.

## RESULTS AND DISCUSSION

### LIBS Spectrum of Infant Formula

In this work, a typical full spectrum and regional potassium peaks of an infant formula are presented in **Figure 2A**. The LIBS spectrum of infant formula has sharp characteristic peaks with different intensities, and each peak uniquely corresponds to a specific element. According to the Atomic Spectra Database (ASD) of National Institute of Standards and Technology (NIST), the peaks located at 766.57 and 769.95 nm were selected for quantifying the potassium content in infant formula. As shown in **Figure 2B**, the spectra of five representative samples with different potassium concentrations were illustrated from 0.415/100 g to 0.815/100 g. It was clear that the intensity of the potassium peaks related to its concentrations accordingly but not linearly, because the potassium peaks were affected by both potassium concentrations and physical parameters (such as laser energy fluctuation and effects related to the sample texture and density). Unfortunately, the contribution of any interference to LIBS was unclear, and DD-LIBS was thus developed to perform the quantitative analysis of potassium by using the existing input-output LIBS data.

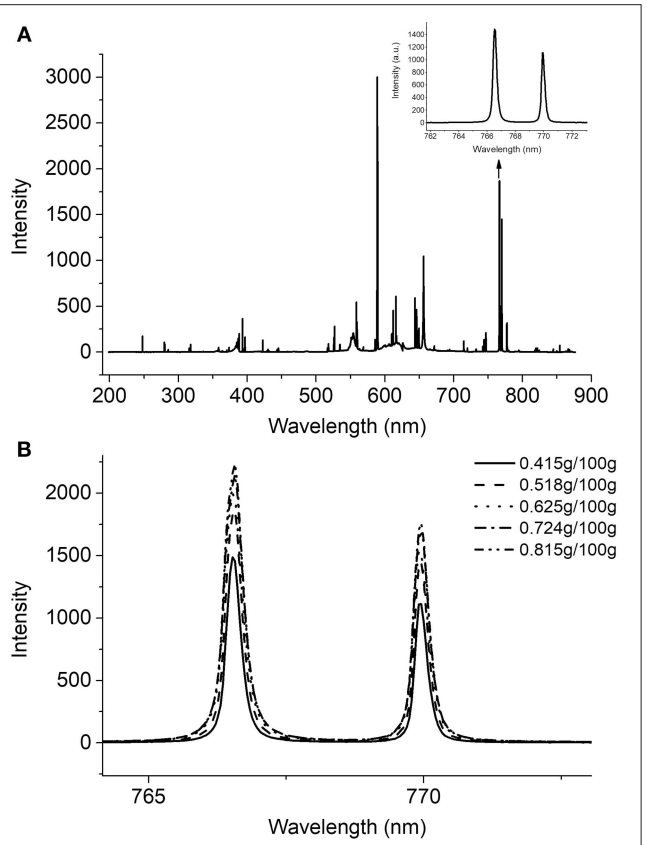



**FIGURE 2 | (A)** A typical LIBS spectrum of infant formula and partially enlarged emission lines of potassium, **(B)** Regional potassium peaks of five samples with different concentrations.

### Selection of Normalization Method

Five normalization methods were compared by calculating the *RMSEP* of each PLS calibration model. The *RMSEPs* of these five normalization methods including average, normalization by norm, spectral area, spectral height, and carbon emission lines, were 0.056, 0.065, 0.076, 0.059, and 0.096, respectively. It is clear that the average normalization strategy was most suitable with the lowest *RSMEP* value and was subsequently applied in this work. After data normalization, the calibration performance of the univariate, PLS and DD-LIBS models was then compared to facilitate the understanding of the LIBS quantification.

### Univariate Analysis

The univariate analysis represents the most conventional modeling strategy, in which the analyte's concentration and the peak intensity or the peak area are set as $x$ and $y$, respectively (El Haddad et al., 2014). In this work, two calibration curves were made with two potassium peaks as shown in **Figures 3A,B**. **Figure 3C** demonstrates another calibration curve using the areas of these two peaks. The LOD obtained from the first peak of potassium was 37 ppm. As shown in **Figures 3A,B**, the $R^2$ of both peak height curves are pretty low, which

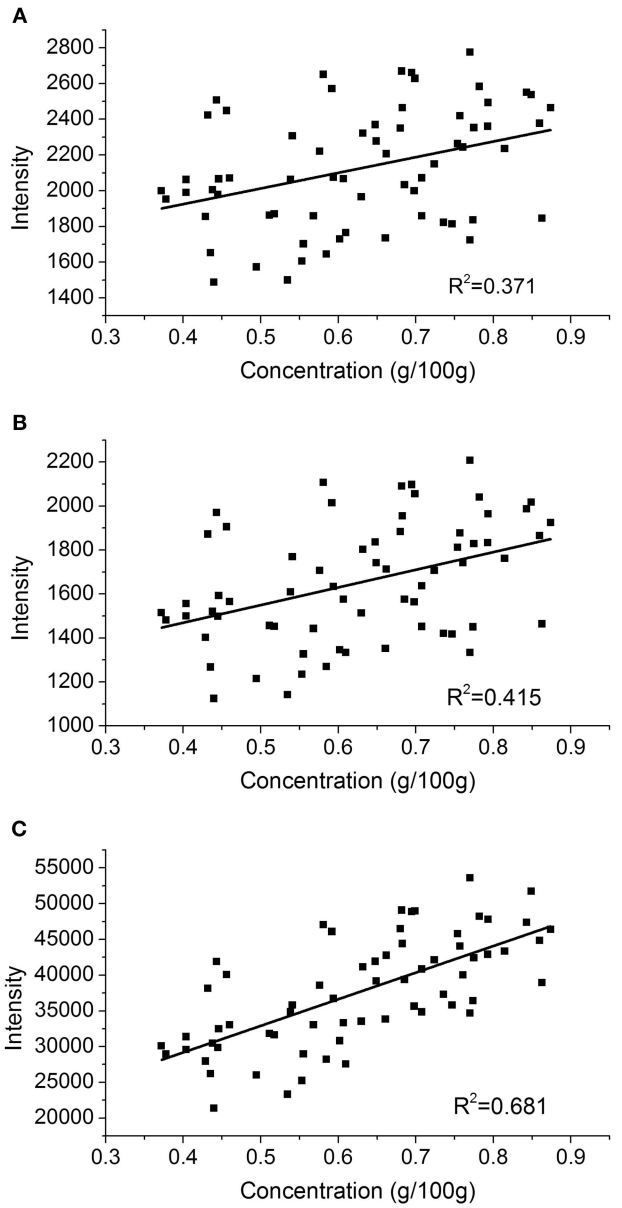



**FIGURE 3 |** A univariate calibration curve based on **(A)** the intensity of the first peak at 766.57 nm, **(B)** the intensity of the second peak at 769.95 nm and **(C)** the areas of two peaks at 766.57 and 769.95 nm.

means that the correlation is poor (El Haddad et al., 2014). The $R^2$ of area (C) is also not satisfactory for quantification even it is slightly higher than the two peaks above-mentioned. The reason is that the univariate analysis is compromised by both matrix effect and sample complexity (Hou et al., 2016; Sanghapi et al., 2016). It is therefore expected that the multivariate analysis could improve the calibration performance through latent projection instead of univariate regression, and PLS was chosen as it is mostly adopted in multivariate calibration.

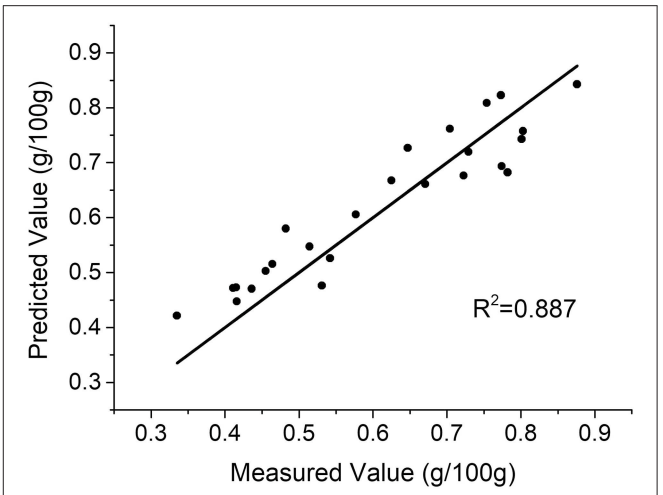



**FIGURE 4 |** Prediction results of the PLS model with the raw LIBS spectra.

## PLS Calibration

The spectral features of potassium were assigned from 751.90 to 774.86 nm, which contains 512 variables. To evaluate prediction capability of the PLS model, $R^2$ and *RMSEP* were calculated. **Figure 4** demonstrates that the prediction results of the PLS model exceed those of univariate analysis. However, the prediction performance could be further improved through the suppression of the uncontrolled spectra shift and broadening.

## DD-LIBS Strategy

In DD-LIBS, the HDWT aims to suppress the effects of peak shift and broadening on multivariate calibration through the oversampling and shift-invariant operation. With the combination of MRFA, DD-LIBS is expected to isolate the spectral features related to the potassium accurately.

### Determination of HDWT Parameters

The performance of HDWT depends on wavelet filters and decomposition scales, which should be optimized before calibration. In HDWT, four wavelet filters with different vanishing moments are available (Selesnick, 2006). Theoretically, the wavelet filter with higher vanishing moment shrinks the peak more efficiently than that with lower vanishing moment (Han et al., 2017). Here, the "bi4" wavelet filter with four vanishing moments was selected, since it possesses the highest vanishing moment in the current HDWT filter bank (Selesnick, 2006). By using the "bi4" filter, the spectral resolution would be expanded by a factor of three, which significantly improved the spectral expandability in an approximately shift-invariant way.

The decomposition scale is also critical in HDWT, so it was optimized by the minimum *RMSECV* criterion. **Figure 5** indicates the relationship between the scale and *RMSECV* using the leave-one-out cross-validation of the calibration set. As a result, the scale four was selected for the HDWT calculation.

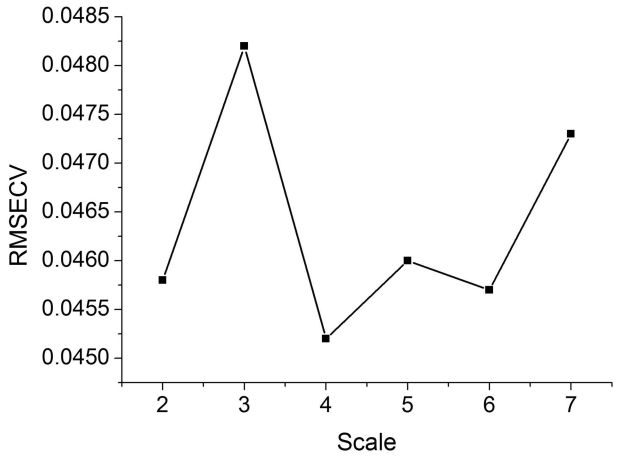



**FIGURE 5 |** RMSECV of the calibration set with different HDWT decomposition scale parameters.

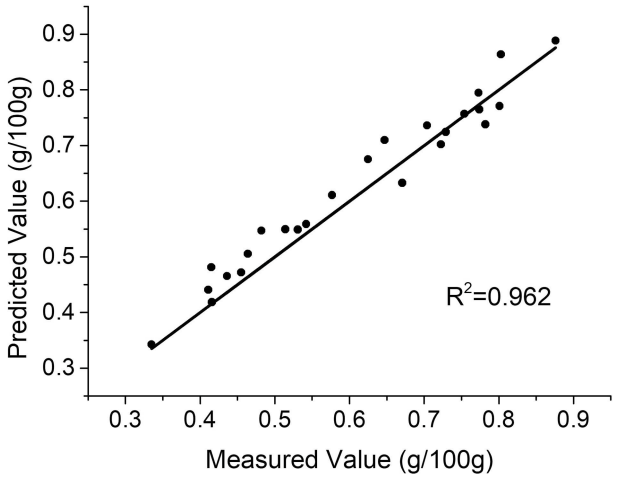



**FIGURE 7 |** Predicted results of the DD-LIBS model.

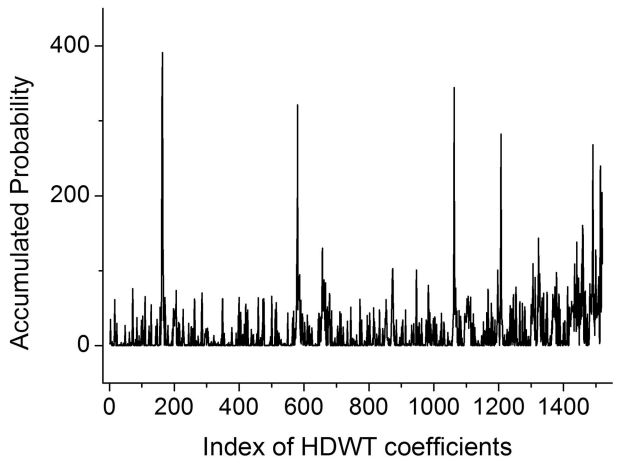



**FIGURE 6 |** Selection probability of HDWT coefficients obtained from the MRFA.

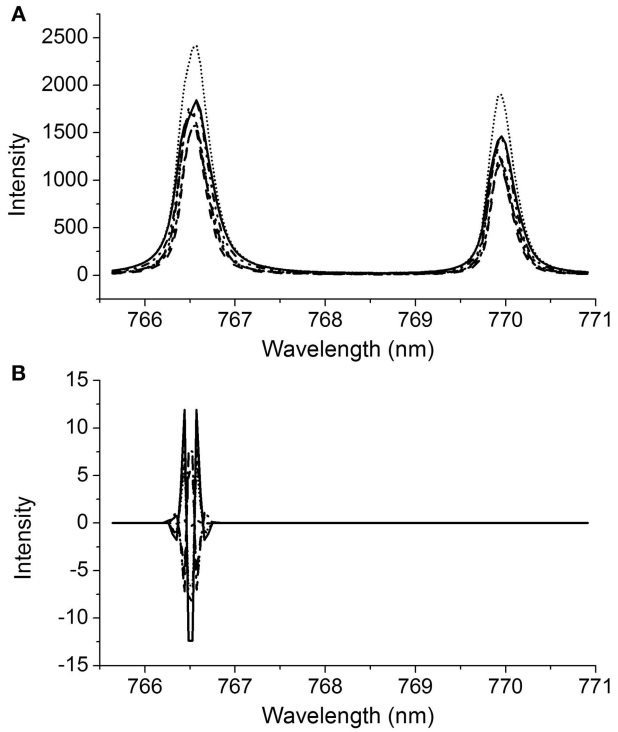



**FIGURE 8 |** LIBS Spectral information obtained from **(A)** potassium and **(B)** DD-LIBS reconstructed spectra.

## Feature Selection Obtained by MRFA

After the HDWT calculation, the original 512 variables were expanded into 1,520 new variables, providing additional flexibility to isolate the features of potassium in the presence of uncontrolled spectral interference. In the sequence, MRFA was adopted to select the accurate features of potassium. **Figure 6** illustrates the accumulated probability of each variable after 1,000 times of MRFA calculation, and the variable with the highest probability was selected for further multivariate calibration.

With the variables selected by MRFA, a PLS model was built. Only one PLS factor was required for calibration, which reveals that DD-LIBS is capable of isolating the spectral peaks of potassium accurately. As compared to **Figure 4**, the $R^2$ of DD-LIBS is improved from 0.887 to 0.962 as shown in **Figure 7**.

It is also of great interest to investigate the reconstructed spectra obtained from the selected variables, which is fundamental to understand how DD-LIBS suppresses the effects of uncontrolled peak shift and broadening on multivariate calibration efficiently. The broadening and shift effect on the LIBS spectral peaks vary from sample to sample as shown in **Figure 8A**, which may impair the LIBS calibration models. As a comparison, the DD-LIBS filtered data is illustrated in **Figure 8B**. It is clear that the reconstructed signals of DD-LIBS locate at

**TABLE 1 |** Prediction results for K content in infant formula.

| Methods | PLS factor | $R^2$ | RMSEP |
|---|---|---|---|
| Univariate (1st peak) | / | 0.099 | 0.423 |
| Univariate (2nd peak) | / | 0.123 | 0.380 |
| Univariate (area) | / | 0.400 | 0.178 |
| PLS | 11 | 0.887 | 0.056 |
| RFA-PLS | 7 | 0.882 | 0.059 |
| HDWT-RFA-PLS | 4 | 0.917 | 0.050 |
| DD-LIBS | 1 | 0.962 | 0.036 |

the same positions as the highest LIBS peak of potassium, and the intensity values at 766.48 and 766.53 nm are the same. It reveals that DD-LIBS cleverly selected the shift-invariant spectral features to overcome the effects of peak shift and peak broadening on multivariate calibration. It is reasonable to expect that DD-LIBS could provide a promising tool to measure potassium content in infant formula accurately, no matter how the uncontrolled interference exists.

## Comparison of Different Methods

**Table 1** shows the prediction results for potassium content in infant formula obtained by different methods. It is obvious that the univariate method presents a poor calibration result, revealing the LIBS spectral analysis should be carefully designed. The PLS model improves the prediction performance of univariate method through multivariate calibration, but the PLS factors are abnormally high. The results illustrate that the additional PLS factors have to be adopted for estimating unknown spectral interference, tending to generate an over-fitting result that relies on the current data set too much. It is unexpected that the combination of RFA and PLS produces a worse result when compared with that of the PLS model. This could be attributed to the effect of spectral interference, e.g., matrix effect, laser energy fluctuation, sample texture and density, and noise, etc. on the feature selection in raw spectra.

The HDWT is explored to suppress the spectral interference. The RFA selects the most important HDWT coefficients, resulting in a better prediction precision than that of the RFA-PLS model. As expected, DD-LIBS provides the best prediction results with only one PLS factor, revealing that the LIBS spectral features of potassium are isolated efficiently. As a result, only one PLS factor is required to construct a high-quality calibration model, thus enhancing the reliability and robustness of the LIBS spectral analysis in the presence of uncontrolled interference.

# CONCLUSION

This study presented a novel strategy, named DD-LIBS, as an approach for real-time quantification of potassium content in commercial infant formula samples. With the combination of HDWT and MRFA, DD-LIBS selected the most important feature related to the potassium accurately, independent of spectral interference. As a result, DD-LIBS generated a high-quality calibration model with only one PLS factor, and the DD-LIBS reconstructed spectra were highly consistent with the original spectral bands of potassium. These satisfactory results suggested a broad expandability of DD-LIBS in the quantification of any targeted element in solid samples in the presence of uncontrolled interference. Once DD-LIBS model has been constructed, it can cleverly predict unknown LIBS spectra as long as these spectra are within a range of relationships learned in the training phase.

# AUTHOR CONTRIBUTIONS

DC planned and supervised the experiments, processed the raw data, revised the manuscript. JZ processed the raw data, wrote the manuscript. JL performed the experiments. ZH advised on data processing and algorithm application. QL revised the manuscript, advised about the principles of LIBS.

# FUNDING

This work was supported by the National Natural Science Foundation of China [61378048, 21305101, 21273159], National Key Research and Development Program of China (2017YFC0803603), Tianjin Research Program of Application Foundation and Advanced Technology [14JCZDJC34700], the Open Funding of State Key Laboratory of Precision Measuring Technology and Instruments [PIL1605], the Program for New Century Excellent Talents in University [NCET-11-0368].

# REFERENCES

Abdel-Salam, Z., Al Sharnoubi, J., and Harith, M. A. (2013). Qualitative evaluation of maternal milk and commercial infant formulas via LIBS. *Talanta* 115, 422–426. doi: 10.1016/j.talanta.2013.06.003

AOAC International (2016). *AOAC SMPR 2014.004 Standard Method Performance Requirements for Minerals and Trace Elements in Infant Formula and Adult/Pediatric Nutritional Formula*.

Aragón, C., and Aguilera, J. A. (2008). Characterization of laser induced plasmas by optical emission spectroscopy: a review of experiments and methods. *Spectrochimi. Acta B At. Spectrosc.* 63, 893–916. doi: 10.1016/j.sab.2008.05.010

Awan, M. A., Ahmed, S. H., Aslam, M. R., Qazi, I. A., and Baig, M. A. (2013). Determination of heavy metals in ambient air particulate matter using laser-induced breakdown spectroscopy. *Arabian J. Sci. Eng.* 38, 1655–1661. doi: 10.1007/s13369-013-0548-7

Bousquet, B., Sirven, J. B., and Canioni, L. (2007). Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples. *Spectrochimi. Acta B At. Spectrosc.* 62, 1582–1589. doi: 10.1016/j.sab.2007.10.018

Cai, W., Li, Y., and Shao, X. (2008). A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemometrics Intell. Lab. Sys.* 90, 188–194. doi: 10.1016/j.chemolab.2007.10.001

Castro, J. P., and Pereirafilho, E. R. (2016). Twelve different types of data normalization for the proposition of classification, univariate and multivariate regression models for the direct analyses of alloys by laser-induced breakdown spectroscopy (LIBS). *J. Anal. At. Spectrom.* 31.2005–2014. doi: 10.1039/C6JA00224B

Chu, X. (2011). *Molecular Spectroscopy Analytical Technology Combined with Chemometrics and its Applications*. Beijing: Chemical Industry Press.

Codex (2015). *Codex Stan 72-1981 Standard for Infant Formula and Formulas for Special Medical Purposes Intended for Infants*. CODEX STAN 72-1981

Cremers, D. A., and Radziemski, L. J. (2013). *Handbook of Laser-Induced Breakdown Spectroscopy, 2nd Edn*. Hoboken, NJ: Wiley.

Deckelbaum, R. J. Adair, L., Appelbaum, M., Baker, G. L., and Baker, S. S. (2004). *Infant Formula: Evaluating the Safety of New Ingredients*. Washington, DC: The National Academies Press.

Dos Santos Augusto, A., Barsanelli, P. L., Pereira, F. M. V., and Pereira-Filho, E. R. (2017). Calibration strategies for the direct determination of Ca, K, and Mg in commercial samples of powdered milk and solid dietary supplements using laser-induced breakdown spectroscopy (LIBS). *Food Res. Int.* 94(Suppl. C), 72–78. doi: 10.1016/j.foodres.2017.01.027

El Haddad, J., Canioni, L., and Bousquet, B. (2014). Good practices in LIBS analysis: review and advices. *Spectrochimi. Acta. B At. Spectrosc*. 101, 171–182. doi: 10.1016/j.sab.2014.08.039

Eseller, K. E., Tripathi, M. M., Yueh, F.-Y., and Singh, J. P. (2010). Elemental analysis of slurry samples with laser induced breakdown spectroscopy. *Appl. Opt*. 49, C21–C26. doi: 10.1364/AO.49.000C21

Gani, A., Gribok, A. V., Rajaraman, S., Ward, K. W., and Reifman, J. (2009). Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling. *IEEE Trans. Biomed. Eng.* 56, 246–254. doi: 10.1109/TBME.2008.2005937

Hahn, D. W., and Omenetto, N. (2010). Laser-induced breakdown spectroscopy (libs), part I: review of basic diagnostics and plasma–particle interactions: still-challenging issues within the analytical plasma community. *Appl. Spectrosc.* 64, 335A−366A. doi: 10.1366/000370210793561691

Hahn, D. W., and Omenetto, N. (2012). Laser-induced breakdown spectroscopy (LIBS), part II: review of instrumental and methodological approaches to material analysis and applications to different fields. *Appl. Spectrosc.* 66, 347–419. doi: 10.1366/11-06574

Han, X., Tan, Z., Huang, Z. X., Chen, X. D., Gong, Y., Li, Q. F., et al. (2017). Nondestructive detection of triclosan in antibacterial hand soaps using digitally labelled Raman spectroscopy. *Anal. Methods* 9, 3720–3726. doi: 10.1039/c7ay00118e

Harmon, R. S., DeLucia, F. C., McManus, C. E., McMillan, N. J., Jenkins, T. F., Walsh, M. E., et al. (2006). Laser-induced breakdown spectroscopy–an emerging chemical sensor technology for real-time field-portable, geochemical, mineralogical, and environmental applications. *Appl. Geochem.* 21, 730–747. doi: 10.1016/j.apgeochem.2006.02.003

Hou, Z., Wang, Z., Yuan, T., Liu, J., Li, Z., and Ni, W. (2016). A hybrid quantification model and its application for coal analysis using laser induced breakdown spectroscopy. *J. Anal. At. Spectromet.* 31, 722–736. doi: 10.1039/C5JA00475F

Hussain, T., and Gondal, M. A. (2008). Detection of toxic metals in waste water from dairy products plant using laser induced breakdown spectroscopy. *Bull. Environ. Contam. Toxicol.* 80: 561. doi: 10.1007/s00128-008-9418-5

ICH Guideline (2005). "Validation of analytical procedures: text and methodology Q2 (R1)," in *International Conference on Harmonization* (Geneva).

ISO (2018a). *Milk Products, Infant Formula and Adult Nutritionals–Determination of Minerals and Trace Elements–Inductively Coupled Plasma Atomic Emission Spectrometry (Icp-Aes) Method.* ISO/DIS 15151 Milk.

ISO (2018b). *Milk Products, Infant Formula and Adult Nutritionals–Determination of Minerals and Trace Elements–Inductively Coupled Plasma Mass Spectrometry (Icp-Ms) Method.* ISO/DIS 21424 Milk.

Kennedy, P. E., and Cade, B. S. (1996). Randomization tests for multiple regression. *Commun. Stat. Simul. Comput.* 25, 923–936. doi: 10.1080/03610919608813350

Koletzko, B., Baker, S., Cleghorn, G., Neto, U. F., Gopalan, S., Hernell, O., et al. (2005). Global standard for the composition of infant formula: recommendations of an ESPGHAN coordinated international expert group. *J. Pediatr. Gastroenterol. Nutr.* 41, 584–599. doi: 10.1097/01.mpg.0000187817.38836.42

Li, H. D., Xu, Q. S., and Liang, Y. Z. (2012). Random frog: an efficient reversible jump markov chain monte carlo-like approach for variable selection with applications to gene selection and disease classification. *Anal. Chim. Acta* 740, 20–26. doi: 10.1016/j.aca.2012.06.031

Li, H. D., Xu, Q. S., and Liang, Y. Z. (2014). libPLS: an integrated library for partial least squares regression and discriminant analysis. *Chemom. Intell. Lab. Syst.* 2018, 34–43. doi: 10.7287/peerj.preprints.190v1

Li, H., Liang, Y., Xu, Q., and Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 648, 77–84. doi: 10.1016/j.aca.2009.06.046

Matsumoto, A., Tamura, A., Koda, R., Fukami, K., Ogata, Y. H., Nishi, N., et al. (2016). A calibration-free approach for on-site multi-element analysis of metal ions in aqueous solutions by electrodeposition-assisted underwater laser-induced breakdown spectroscopy. *Spectrochimi. Acta B At. Spectrosc.* 118(Suppl. C), 45–55. doi: 10.1016/j.sab.2016.02.005

Meucci, V., Razzuoli, E., Soldani, G., and Massart, F. (2010). Mycotoxin detection in infant formula milks in Italy. *Food Addit. Contam. A* 27, 64–71. doi: 10.1080/02652030903207201

Moros, J., Kuligowski, J., Quintás, G., Garrigues, S., and de la Guardia, M. (2008). New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs. *Anal. Chim. Acta* 630, 150–160. doi: 10.1016/j.aca.2008.10.024

Panne, U., Neuhauser, R. E., Theisen, M., Fink, H., and Niessner, R. (2001). Analysis of heavy metal aerosols on filters by laser-induced plasma spectroscopy. *Spectrochimi. Acta B At. Spectrosc.* 56, 839–850. doi: 10.1016/S0584-8547(01)00209-9

Parab, G. S., Rao, R., Lakshminarayanan, S., Bing, Y. P., Moochhala, S. M., and Swarup, S. (2009). Data-driven optimization of metabolomics methods using rat liver samples. *Anal. Chem.* 81, 1315–1323. doi: 10.1021/ac801645t

Qin, Y., Tang, B., and Wang, J. (2010). Higher-density dyadic wavelet transform and its application. *Mech. Syst. Signal Process.* 24, 823–834. doi: 10.1016/j.ymssp.2009.10.017

Sanghapi, H. K., Jain, J., Bol'shakov, A., Lopano, C., McIntyre, D., and Russo, R. (2016). Determination of elemental composition of shale rocks by laser induced breakdown spectroscopy. *Spectrochim. Acta Part B Atom. Spectrosc.* 122, 9–14. doi: 10.1016/j.sab.2016.05.011

Selesnick, I. W. (2006). A higher density discrete wavelet transform. *IEEE Trans. Signal Process.* 54, 3039–3048. doi: 10.1109/TSP.2006.875388

Soetan, K., Olaiya, C., and Oyewole, O. (2010). The importance of mineral elements for humans, domestic animals and plants-a review. *Afr. J. Food Sci.* 4, 200–222.

Sun, L., and Yu, H. (2009). Correction of self-absorption effect in calibration-free laser-induced breakdown spectroscopy by an internal reference method. *Talanta* 79, 388–395. doi: 10.1016/j.talanta.2009.03.066

The Ministry of Health People's Republic of China (2010a). *National Food Safety Standard Determination of Calcium, Iron, Zinc, Sodium, Potassium, Magnesium, Copper and Manganese in Foods for Infants and Young Children, Raw Milk and Dairy Products.* GB 5413.21–2010.

The Ministry of Health People's Republic of China (2010b). *National Food Safety Standard Infant Formula.* GB 10765-2010.

Tan, Z., Lou, T. T., Huang, Z. X., Zong, J., Xu, K. X., Li, Q. F., et al. (2017). Single-drop raman imaging exposes the trace contaminants in milk. *J. Agri. Food Chem.* 65, 6274–6281. doi: 10.1021/acs.jafc.7b01814

Yun, Y. H., Li, H. D. E., Wood, L. R., Fan, W., Wang, J. J., Cao, D. S., et al. (2013). An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration. *Spectrochimi. Acta A. Mole. Biomol. Spectros.* 111(Suppl. C), 31–36. doi: 10.1016/j.saa.2013.03.083

Zheng, J., Lu, J., Zhang, B., Dong, M., Yao, S., Lu, W., et al. (2014). Experimental study of laser-induced breakdown spectroscopy (libs) for direct analysis of coal particle flow. *Appl. Spectros.* 68, 672–679. doi: 10.1366/13-07278

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 28 August 2018
doi: 10.3389/fchem.2018.00363

# Essentials of Aquaphotomics and Its Chemometrics Approaches

*Roumiana Tsenkova[1]*, Jelena Munćan[1,2], Bernhard Pollner[3] and Zoltan Kovacs[4]*

[1] *Biomeasurement Technology Laboratory, Graduate School of Agricultural Science, Kobe University, Kobe, Japan,* [2] *Nanolab, Biomedical Engineering Department, Faculty of Mechanical Engineering, University of Belgrade, Belgrade, Serbia,* [3] *Department for Hygiene and Medical Microbiology, Medical University of Innsbruck, Innsbruck, Austria,* [4] *Department of Physics and Control, Faculty of Food Science, Szent István University, Budapest, Hungary*

Aquaphotomics is a novel scientific discipline involving the study of water and aqueous systems. Using light-water interaction, it aims to extract information about the structure of water, composed of many different water molecular conformations using their absorbance bands. In aquaphotomics analysis, specific water structures (presented as water absorbance patterns) are related to their resulting functions in the aqueous systems studied, thereby building an aquaphotome—a database of water absorbance bands and patterns correlating specific water structures to their specific functions. Light-water interaction spectroscopic methods produce complex multidimensional spectral data, which require data processing and analysis to extract hidden information about the structure of water presented by its absorbance bands. The process of extracting information from water spectra in aquaphotomics requires a field–specific approach. It starts with an appropriate experimental design and execution to ensure high-quality spectral signals, followed by a multitude of spectral analysis, preprocessing and chemometrics methods to remove unwanted influences and extract water absorbance spectral pattern related to the perturbation of interest through the identification of activated water absorbance bands found among the common, consistently repeating and highly influential variables in all analytical models. The objective of this paper is to introduce the field of aquaphotomics and describe aquaphotomics multivariate analysis methodology developed during the last decade. Through a worked-out example of analysis of potassium chloride solutions supported by similar approaches from the existing aquaphotomics literature, the provided instruction should give enough information about aquaphotomics analysis i.e. to design and perform the experiment and data analysis as well as to represent water absorbance spectral pattern using various forms of aquagrams—specifically designed aquaphotomics graphs. The explained methodology is derived from analysis of near infrared spectral data of aqueous systems and will offer a useful and new tool for extracting data from informationally rich water spectra in any region. It is the hope of the authors that with this new tool at the disposal of scientists and chemometricians, pharmaceutical and biomedical spectroscopy will substantially progress beyond its state-of-the-art applications.

**Keywords: aquaphotomics, water, near infrared spectroscopy, multivariate analysis, water spectral pattern, aquagram, aquap2**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Daniel Cozzolino,*
*Central Queensland University, Australia*
*Felix Scholkmann,*
*UniversitätsSpital Zürich, Switzerland*
*Zhisheng Wu,*
*Beijing University of Chinese Medicine, China*

****Correspondence:***
*Roumiana Tsenkova*
*rtsen@kobe-u.ac.jp*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *15 April 2018*
***Accepted:*** *30 July 2018*
***Published:*** *28 August 2018*

***Citation:***
*Tsenkova R, Munćan J, Pollner B and Kovacs Z (2018) Essentials of Aquaphotomics and Its Chemometrics Approaches. Front. Chem. 6:363. doi: 10.3389/fchem.2018.00363*

## INTRODUCTION TO AQUAPHOTOMICS

Aquaphotomics is a novel scientific discipline founded by Professor Roumiana Tsenkova at Kobe University, Japan, in 2005 (Tsenkova, 2005, 2006a,b,c, 2009) with the objective of studying and systematizing knowledge about water-light interaction, which was found to be a huge source of information on the subject of the structural and related functional properties of aqueous systems. This is a complementary "omics" discipline dealing with the large-scale, comprehensive study of water as the "*molecular and energy mirror*" of the rest of the aqueous system. While proteomics studies proteins, glycomics—carbohydrates and lipidomics—lipids; aquaphotomics explores the roles, relationships and functions of the water—an equally important biomolecule and one of nature's fundamental building blocks.

The word "aquaphotomics" is derived from the words *aqua*—water and *photo*-light since this new discipline studies water by using its interaction with the light. Thus, aquaphotomics is a science which uses water-light interaction to explore the structure of water—as a system and matrix composed of many different water molecular conformations, thereby resulting in various functionalities (Tsenkova, 2009). The main objective of establishing aquaphotomics as a novel scientific discipline was to provide a common platform and strategy to lead to an improved general understanding of the water functionality by utilizing water-light interaction at every frequency of the electromagnetic spectrum. The majority of aquaphotomics works so far have been done by using near infrared (NIR) spectroscopy, especially in the area of the 1st overtone of the OH stretching band (1,300–1,600 nm) where many water absorbance bands are identified and consistent with previously reported or calculated overtones of water absorbance bands in the infrared region (Weber et al., 2000, 2001; Smith et al., 2005; Tsenkova, 2009; Tsenkova et al., 2015). What aquaphotomics research studies showed is that NIR spectroscopy, and in general water-light interaction over the entire electromagnetic spectrum, can significantly contribute to the field of water science and better understanding of water molecular systems (Tsenkova, 2009).

The NIR wavelength region from around 680 to 2,500 nm is considered as an excellent tool for water observation that provides an enormous amount of information about water molecular structure (Büning-Pfaue, 2003; Tsenkova, 2009). The NIR light allows a longer penetration length, as compared to infrared, even up to 10 mm in the short wavelength region (750–1,100 nm) (Workman, 2000), making it a rapid and non-destructive measurement technique particularly suitable for studying intact biological systems. Numerous NIR spectra can be obtained in various conditions and states of the systems (under different perturbations)—all in real time. NIR spectroscopy has a rich history of applications in pharmaceutical and medical fields. Water, however, with its NIR characteristic spectrum was often seen as a problematic component and the common source of measurement error, because it could alter sample spectra, hide weak absorbance bands and shift other absorbance bands (Ciurczak and Igne, 2014). In fact, water is cited as one of the main disadvantages of NIR spectroscopy in pharmaceutical applications since it prevents a direct quantification (Jamrógiewicz, 2012).

Traditionally, water bands in the NIR region around 1,440 nm (the first overtone of OH stretch) and 1,940 nm (a combination of OH bending and stretching) have been very useful in the studies of the state of water in various samples (Ozaki, 2002). One of the major and most common applications of NIR spectroscopy was moisture determination (Osborne et al., 1993; Reeves, 1995). NIR spectroscopy has been used to investigate water content, hydrogen bonds and hydration state in a variety of fields such as agriculture and food industry, medical and pharmaceutical sciences, and polymer and textile industries (Ozaki, 2002).

Although some early works on water analysis reported the rich informational potential of its NIR spectrum (Hirschfeld, 1985; Iwamoto et al., 1987; Grant et al., 1989; Maeda et al., 1995), it was only with the development of aquaphotomics that the properties of water as a "collective matter and energy mirror" were truly explored (Tsenkova, 2009). The so-called "*water mirror approach*" of aquaphotomics utilizes the high sensitivity of water's hydrogen bonds, where all the components of the aqueous system and surrounding energies influence the water structure, i.e., the covalent bonds. Every aqueous system is a dynamic arrangement of water molecular network hydrogen-bonded to other constituents and influenced by perturbations. Any perturbation of the aqueous system results in changes of water molecular conformations, which in turn produce changes in the corresponding NIR spectra at their respective water absorbance bands. As a consequence of the strong potential of water molecules for hydrogen bonding, water, a natural matrix of any aqueous or biological system, changes its absorbance pattern every time it adapts to a physical or chemical change in the system itself or its environment (Tsenkova, 2008c). It is this quality of water that indirectly permits measurements of small quantities or structural changes of other molecules present in the aqueous system. By tracking the changes of water absorbance bands in the spectra of aqueous or biological systems, the information is extracted about not only water structure but also other components present in water or the state of the system as a whole (Tsenkova, 2006c, 2007, 2008b, 2009).

Being rapid and non-destructive, NIR spectroscopy is a powerful technique with an incredible range of applications, whose horizons have been further expanded by aquaphotomics. Since its establishment more than a decade ago, aquaphotomics has grown into a vast and multidisciplinary scientific field, encompassing many research areas (**Table 1**). Changes in the absorption spectrum of water are used for quantification of the solutes present in water, even when the solutes do not absorb NIR light at all (Grant et al., 1989; Tsenkova, 2009; Gowen et al., 2015). This so-called water-mirror approach enables measurements of concentrations previously impossible with NIR spectroscopy at ppm levels (Sakudo et al., 2006b; Tsenkova, 2008b; Gowen et al., 2013; Bázár et al., 2014, 2015), and even at ppb levels under certain experimental conditions (Sakudo et al., 2005, 2006b; Tsenkova et al., 2007b; Tsenkova, 2008a,b). Furthermore, the aquaphotomics research of biological systems introduced a concept of water spectral pattern as a holistic biomarker (Tsenkova, 2006c, 2007), which relates

certain structures of water with functionalities of the respective biological systems, thus opening new directions toward non-destructive quality monitoring applications and non-invasive biodiagnosis.

The aquaphotomics research fields have two things in common. First, water is the common matrix of all the systems studied. Second, the approach to extract the information hidden in complex and multidimensional spectra of such systems requires a specific aquaphotomics methodology developed over the years and based on rich experience in dealing with a great variety of aqueous systems. The objective of this paper is to provide guidance about how to perform aquaphotomics analysis of NIR data. Using an example dataset of aqueous salt solutions, each step of the analysis will be explained and supplemented by similar examples from the existing literature illustrating how specific steps in data analysis provide new insights, improve spectral quality, or reveal new information. The basic methodology explained in this work is applicable to the analysis of NIR data of any aqueous system, with minor aqueous system- and purpose-specific adjustments. A step-by-step explanation of aquaphotomics analysis supplemented by citations of similar works will provide a solid basic knowledge about how to start and perform the analysis as well as where to look for further information. It is the hope of the authors that, with this new tool at the disposal of scientists and chemometricians, pharmaceutical and biomedical spectroscopy will utilize the richness of NIR water spectra to extend its applications far beyond moisture determination, leading to a substantial progress beyond the current state of the art.

## GLOSSARY OF AQUAPHOTOMICS TERMS

This glossary is intended to define the terms and certain abbreviations commonly used in the aquaphotomics literature, which will appear throughout this paper. New terminology has emerged over time and with the development of aquaphotomics and the resulting need to better describe its subject of exploration using newly discovered knowledge. The origin and definitions for the terms are compiled from several sources, which are listed in the respective columns of **Table 2**.

With the main terms explained, we can now formulate the objective of aquaphotomics analysis i.e., the water mirror approach to analyze aqueous systems as a whole, using their multidimensional spectra and focusing on water absorbance bands located at specific regions, allows observation and absorbance measurements. When activated water absorbance bands are found in response to some perturbation of interest, then a water absorbance spectral pattern caused by the respective perturbation is identified. By compiling water absorbance patterns in an aquaphotome, aquaphotomics builds up a comprehensive database of the states of the analyzed system as a whole, in terms of identified water structures shaped by various internal or external perturbations. In future applications, aquaphotome database will provide a rapid identification of causes for changes and influences on the system based on the recognized water spectral patterns, which serve as holistic markers of the state of the aqueous system or biomarkers in the case of biological systems (Tsenkova, 2006c; Kovacs et al., 2016).

## AQUAPHOTOMICS METHODS

### Basic Workflow and General Guidance

The basic workflow of aquaphotomics analysis from the experimental design to the final act of building an aquaphotome is illustrated in **Figure 1**. Similar to every conventional NIR spectroscopy work, everything starts with a proper experimental design and instrumental setup.

Although NIR spectroscopy, in general, does not require sample preparation, there are some specific aspects in aquaphotomics experimental design requiring more attention.

First of all, it is an absolute must to ensure that the instruments have high-quality spectral signals. In general, not all spectrometer systems are suited for aquaphotomics experiments. It is advisable to check the instrument's performance beforehand to ensure the high quality of the spectra in the entire Vis-NIR region (400–2,500 nm). All subsequent analysis will be highly influenced by the quality of raw spectral data. It is therefore of the utmost importance to evaluate raw spectra prior to any real experimental work. The basic analytical procedures for detecting errors of NIR data and evaluation of signal quality have been recently provided in an extensive study performed by Bazar et al., which tested and compared the performance of three spectrometer systems (Bazar et al., 2016). This paper can be used as a general guidance on how to test the quality and performance of NIR instrument before venturing further.

Ensuring good spectral quality is particularly important since, in addition to the already known complexity of NIR spectra due to the overtone and combination modes resulting in broad bands, the changes in the spectra of aqueous systems caused by some perturbation of interest are small and subtle. The useful information may end up being buried in noise if the instrument does not provide a high signal-to-noise ratio. Another prerequisite is the use of a high-resolution instrument. Water absorbance bands in the NIR range are usually located very close to each other, so high spectral resolution of 0.5 or 1 nm will ensure an optimal detection and separation of the bands in a subsequent analysis.

An experiment should be carried out according to previously defined protocols to ensure the same environmental conditions. The purpose of carefully designed and established protocols is to minimize the influence of unknown factors that may affect sample spectra.

The specificity of experimental design may vary depending on the type of aqueous system involved; however, the design must ensure that each sample is presented with several replicates (sample replicates) and each measurement is performed by using several consecutive illuminations (consecutive replicates, consecutive spectra). Collecting and averaging multiple scans is part of the standard practice to remove noise—recoding 64 or more scans per one spectrum reduces the noise levels significantly (Manley, 2014). Measuring liquid samples should always start with pure water (18.2 MΩ·cm) and all subsequent measurements should be done with a cuvette always placed in

**TABLE 1 |** Fields of aquaphotomics applications.

| Application | References |
|---|---|
| Fundamental biochemical studies of water solutions | Sugars (Bázár et al., 2015; Cui et al., 2017a), proteins (Tsenkova et al., 2004; Chatani et al., 2014), DNA (Goto et al., 2015), salts (Gowen et al., 2013, 2015), alkali-metal halides (Kojić et al., 2014), acids (Omar et al., 2012), and metal ions (Sakudo et al., 2006b; Tsenkova et al., 2007a; Putra et al., 2010) |
| Water quality | Water filtration process (Cattaneo et al., 2011), detection and quantification of pesticides (Gowen et al., 2011), discrimination of mineral waters (Munćan et al., 2014), detection of contaminants (Gowen et al., 2015), and holistic water monitoring (Kovacs et al., 2016) |
| Food quality | Various foodstuff (Gowen, 2012), cheese (Atanassova, 2015), honey (Bázár et al., 2016), mushrooms (Gowen et al., 2009a), bacteria in food (Nakakimura et al., 2012), milk (Tsenkova, 1994; Tsenkova et al., 2001a,b), and food packaging influence (Cattaneo et al., 2016; Barzaghi et al., 2017) |
| Materials and nanomaterials | Soft contact lenses (Munćan et al., 2016b; Šakota Rosić et al., 2016) fullerene based nanomaterials (Matija et al., 2012, 2017), and polystyrene particles (Tsenkova et al., 2007b) |
| Microbiology | Bacteria (Nakakimura et al., 2012; Remagni et al., 2013; Slavchev et al., 2015, 2017),and HIV virus (Sakudo et al., 2005) |
| Plant biology | Mosaic virus detection in soybeans (Jinendra et al., 2010), and abiotic and biotic stress (Jinendra, 2011) |
| Animal medicine | Mastitis in cows (Tsenkova et al., 2001a,b,c, 2005; Tsenkova and Atanassova, 2002; Atanassova et al., 2009; Meilina et al., 2009), udder health (Tsenkova, 1994), ovulation period in Bornean orangutan (Kinoshita et al., 2016), ovulation period in giant pandas (Kinoshita et al., 2010, 2012), estrus detection in cows (Takemura et al., 2015), and tissue discrimination (Sakudo et al., 2006a) |
| Human medicine | DNA mutations (Goto et al., 2015), HIV virus detection (Sakudo et al., 2005), tissue discrimination (Sakudo et al., 2006a), the state of metals in tissues (Sakudo et al., 2007), prion protein disease (Tsenkova et al., 2004), skin cream effects (Matija et al., 2013) dialysis efficacy monitoring (Munćan et al., 2016a), colorectal cancer diagnostics (Munćan et al., 2016a) |

the same position (the same side). The same cuvette should be used throughout the experiment. It should be first rinsed at least in triplicate with sample before final filling. After that, it is placed in the sample holder and allowed to equilibrate before scanning in order to minimize inter-sample variation.

Reference measurement (blank air) should be done before each sample measurement. The order of sample measurement and sample replicates should be completely randomized; but pure water should be always scanned after a previously defined number of samples (e.g., every 5, 7, or 10 sample measurements). There are two reasons for measurements of pure water in between samples. First, these spectra are used as an environmental control, monitoring known and unknown influences on water and could later be used to correct or remove unwanted influences from sample spectra. Second, it builds a large library of pure water spectra. There are many advantages of building such a library—it contains the spectra of pure water under various changing conditions over a longer period of time under different temperatures, humidity conditions and various day-to-day variations of the instrument and working environment. Building such a database has been proved very useful for correction in general NIR applications (Tillmann and Paul, 1998). In addition, a novel method for enhancement of spectral signals has been recently developed, which also relies on building a similar library (Kojić et al., 2017).

It is also advisable to monitor and log major external influences such as laboratory temperature, atmospheric pressure and humidity, as well as sample holder temperature or cuvette. Measuring and logging external parameters can be very useful for identification of major sources of spectral variation as well as for exploration of the dynamics of different aqueous systems under the same environmental perturbations.

As opposed to traditional NIR spectroscopy, which places emphasis on the control of the environment during the measurements, "perturbation" is often used in aquaphotomics and is sometimes even a necessary component of experiments, which helps in revealing hidden information. The analysis of aqueous systems' spectra under the influence of some chosen, intentional, perturbation can be defined as an evaluation of the system by applying changes to the selected parameters and re-estimation of the results (Tsenkova, 2007). In practice, the most frequently used perturbations to induce changes in the respective systems are changes in temperature (Gowen et al., 2013; Chatani et al., 2014; Putra et al., 2017; Wenz, 2018), consecutive illuminations (Tsenkova, 2005; Chatani et al., 2014; Wenz, 2018), and changes in dilution (Gowen et al., 2013; Wenz, 2018). Other types of perturbations can also be used to test the robustness of the models developed. Besides temperature perturbation, for example Putra et al. (2017) and Meilina et al. (2011) introduced perturbations by different metal ions to test the regression model developed for the measurement of cadmium concentrations in aqueous solutions. The use of intentional, artificially created perturbations provides a change in entropy and leads to the revelation of hidden spectral information (Tsenkova, 2006c). A recent work by Wentz on water in model membranes employed four types of perturbation in the same work in order to probe and thoroughly examine changes in the water matrix [i.e., temperature, consecutive illuminations, concentration (dilution)], and difference in molecular structure of phospholipids (fourteen identical carbon acyl chains but with polar heads differing in the presence of an hydroxyl or a choline group) (Wenz, 2018). The most frequently used intentional perturbations (consecutive illuminations or increasing temperature) result in similar changes in water matrix—an increase in the number of free water molecules, which are then available for "scanning" of the rest of the system; in other words—to interact with its components, which results in changes in sample spectra and provision of additional

**TABLE 2 |** Glossary of aquaphotomics terms.

| Term | Definition |
|---|---|
| Water Mirror Approach (Tsenkova, 2008b, 2009) | Aquaphotomics spectral analysis is often called "water mirror approach" because of the indirect manner of acquiring information about solute composition or surroundings of the aqueous system, namely by measuring the changes in absorbance at water absorbance bands in the spectrum of the aqueous system (Tsenkova, 2009). |
| WAMACS - Water Matrix Absorbance Coordinates (Tsenkova, 2009) | The WAMACS are spectral ranges, where specific water absorbance bands related to specific water molecular conformations (water species, water molecular structures) are found with the highest probability (Tsenkova, 2009). For the first overtone of water (1300-1600nm), 12 WAMACs (labeled Ci, i=1, 12) have been experimentally discovered (each 6-20nm width) and they have been confirmed by overtone calculations of already reported water bands in the infrared range (Tsenkova, 2009). |
| WABS – Water Absorbance Bands (Tsenkova, 2009) | Studies in the infrared range have identified the absorbance bands of numerous water species (Buijs and Choppin, 1963; Fornés and Chaussidon, 1978; Doster et al., 1986; Maeda et al., 1995; Sartor et al., 1995; Luck, 1998; Czarnik-Matusewicz et al., 1999; Heiman and Licht, 1999; Murayama et al., 2000; Segtnan et al., 2001; Chandler, 2002; Cupane et al., 2002; Šašić et al., 2002; Robertson et al., 2003). When their overtones are calculated, it is confirmed that together with already known bands, these bands occur within the whole Vis-NIR range (Tsenkova, 2005). So far, the spectral database of water absorbance bands has more than 500 bands in the area of the first, second and third overtones of water (Tsenkova, 2009; Tsenkova et al., 2015). The systematization of already identified and discovery of new water absorbance bands related to specific water species structures is one of the ongoing aquaphotomics endeavors. |
| Activated water bands | When a certain perturbation of interest is shown to produce the changes at specific water absorbance bands, and when this is determined consistently and repeatedly throughout the aquaphotomics analysis, these water absorbance bands are considered "activated" by the respective perturbation. |
| WASP–Water Absorbance Spectral Pattern (Tsenkova, 2009) | The combination of the *activated water bands* caused by a certain perturbation defines water absorbance spectral pattern, which describes the condition of the whole aqueous system. WASP can contain huge amounts of chemical and physical information about the respective aqueous system and can be thought of as a holistic marker because it captures the structure and dynamics of the respective system as a whole. At the moment, even without the assignment and understanding of water absorbance bands, WASPs can be used as holistic (bio) markers for system functionality. |
| Aquagrams (Tsenkova, 2010) | An aquagram is a novel graphical representation of data, invented to present in a succinct manner a water absorbance spectral pattern – WASP (Tsenkova, 2010). |
| Aquaphotomes (Tsenkova, 2009) | An aquaphotome is the entire complement of water molecular structures produced by aqueous or biological systems in different conditions. It can be defined as a comprehensive database of all water spectral patterns with the interpretation of their functionality given a particular set of conditions of the respective system, (Tsenkova, 2009). Every aquaphotome is system-specific. Once a large database of characteristic water bands has been acquired, they can be related to specific biological functions and subsequently used for prediction, diagnosis, and understanding of biology, chemistry and physics of biological and aqueous systems (Tsenkova, 2009). |

information. Regarding unintentional perturbations, it is always advisable to investigate what perturbations (i.e., factors) have an influence on the developed models. These perturbations may include individual differences or the presence of disease in the case of biological systems studied, or even sample thickness (Tsenkova, 2004).

The first step of analysis begins with the inspection of raw spectral data. Although NIR spectra of aqueous systems are comprised of broad, overlapping spectral bands, visual spectral inspection still remains a vital step before any further data analysis. Visual inspection gives the first clues about the presence of outliers, helps in deciding what preprocessing steps to proceed with, gains a general insight into how samples are grouped and on what spectral regions to focus the attention. All the subsequent steps—data preprocessing, conventional spectral analysis and chemometrics application, which will be described in more detail later-serve to extract the information of interest. From the aspect of conventional data analysis—with building, testing and validation of a model—either qualitative or quantitative, depending on the objective of the experiment, the work is done when suitable prediction accuracy is achieved. However, this is only half of the work done in an aquaphotomics analysis. Each step of the analysis—raw data inspection, preprocessing, conventional and chemometrics analysis (an array of exploratory, classification and regression analysis)—provide certain quantitative outputs like derivatives, subtracted spectra, regression vectors or loading vectors, discriminating power and others, which all unravel water absorbance bands most affected by perturbation of interest (WABS, **Figure 1**).

The NIR spectra of aqueous systems are very complex, and changes in their absorbance spectra caused by some perturbation will usually be very subtle, but nonetheless persistent and consistent. From all the WABs discovered during multiple steps of aquaphotomics analysis, a noticeable pattern of repeating, common absorbance bands will emerge to reveal perturbation-induced water absorbance bands i.e., how and what water molecular conformations are affected. When this absorbance spectral pattern water absorbance pattern (WASP) is recognized, it can be presented in a simple, yet concise and informative manner by using aquagrams. This aspect of aquaphotomics analysis adds one more dimension to the results obtained in that it provides understanding of the water functionality in the respective system. It allows linking discovered WASPs with the conditions of the aqueous systems analyzed, revealing how and why water changes the way it does under certain perturbation. This is of special importance for living, biological systems. The storing of WASPs into a large aquaphotome database allows for a fast comparison and identification of the state of aqueous or biological systems, thereby in essence providing biodiagnosis based on the state of water.

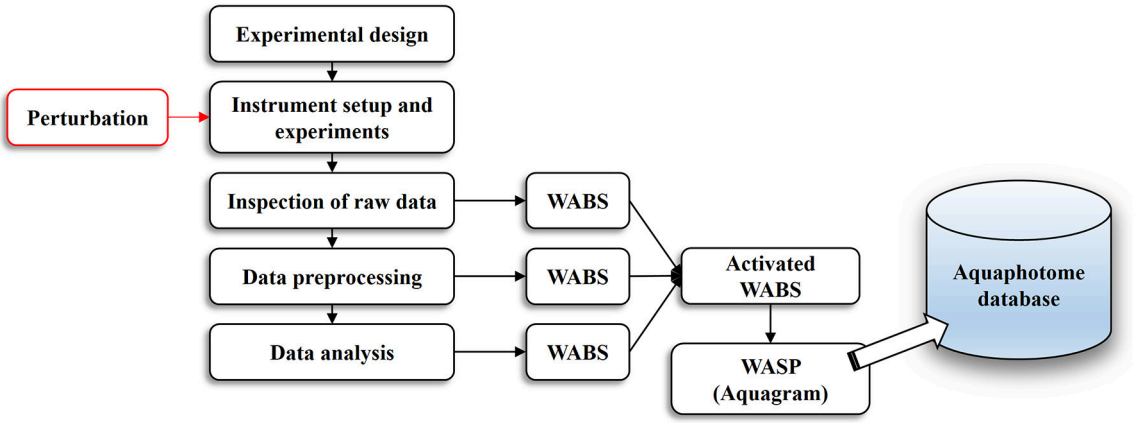



**FIGURE 1 |** An overview of the aquaphotomics basic methodology for design, performance and analysis of experimental data with the aim of extracting water spectral pattern for the defined perturbation.

## Aquaphotomics Analysis of Potassium Chloride Solutions—A Worked-Out Example

To better illustrate the working process of aquaphotomics analysis, we will present an example of analysis performed on the spectral dataset of aqueous solutions of potassium chloride in the next sessions. The perturbation of the water matrix by salt and measurement of salt concentration are already available in aquaphotomics literature (Gowen et al., 2015) and even in very early near infrared spectroscopy applications (Grant et al., 1989). We have chosen this perturbation since it perfectly illustrates the aquaphotomics water-molecular and energy mirror concept in that the salts are practically transparent for NIR light. Therefore, the results obtained thereby are based entirely on the changes in the water molecular matrix. Experimental condition will be described next.

### Materials and Methods

#### *Sample preparation*

Potassium-chloride (KCl, M = 74.56 g.mol$^{-1}$, purity ≥ 99.0% w/w, Wako Pure Chemical Industries, Ltd. Kobe, Japan) was used.

All samples were prepared by using deionized water from a Milli-Q water purification system (Millipore, Molsheim, France). A stock solution of 100 mM was prepared at first. Working solutions were made by serial dilution of the stock solution in 10-mM steps to produce the following KCl concentrations: 10, 20, 30, 40, 50, 60, 70, 80, and 90 mM. All samples of the stock and working solutions were freshly prepared in two independent sample replicates (i.e. a total of 20 samples for the analysis).

#### *NIR spectra collection*

Transmittance spectra of KCl aqueous solutions were acquired by using a FOSS-XDS spectrometer (FOSS NIRSystems, Inc., Hoganas, Sweden) equipped with a Rapid Liquid Analyzer module consisting of a temperature-controlled cuvette holder. The temperature of the sample holder was kept constant at 28°C during all measurements. This temperature was chosen to be close to the ambient temperature (ca. 28°C), allowing a fast and easy way of maintaining constant temperature during measurements. Each sample was firstly incubated in the sample holder for 90 s before scanning to get the required temperature of 28°C. Deionized water samples were measured as an environmental control for every five sample measurements. Spectral acquisition order was randomized with respect to salt concentration. The 1-mm path length quartz sample cell was used as a container.

The spectra were acquired in the range of 400–2,500 nm, with a resolution of 0.5 nm. Each saved spectrum was an average of 32 successive scans. This number of scans was chosen to shorten the acquisition time. Three consecutive spectra were recorded for each sample and for each measurement. The reference spectrum was recorded before each measurement. The spectral data were transformed to pseudo-absorbance units (log$T^{-1}$, where $T$ = transmittance). One sample was represented by six spectra in total, from two independent sample replicates and three consecutive spectra. The total number of recorded spectra was 75 (10 concentrations × 2 sample replicates × 3 consecutive scans + 15 control scans of deionized water).

The FOSS-XDS instrument was operated by using VISION 3.5 software (FOSS NIRSystems, Inc., Hoganas, Sweden).

#### *Data analysis*

For the purpose of this paper, the data analysis of KCl solutions was performed by using only the wavelength range from 1,300 to 1,600 nm, which represents the absorption region of OH bonds of water (1st overtone of OH).

Smoothed spectra were calculated by using a Savitzky-Golay polynomial filter (2nd order polynomial fit and 21 points). Difference spectra were calculated by subtraction the average spectrum of deionized water from the average spectra of potassium-chloride solutions for each concentration level. The 2nd derivative spectra of potassium-chloride solutions were calculated by using a Savitzky-Golay filter (2nd order

polynomial fit and 21 points). Principal component analysis (PCA) was used to describe multidimensional patterns in the spectral data and to discover outliers. The relationship between the actual and predicted concentrations of KCl was examined by using Partial Least Squares Regression (PLSR) based on leave-one (concentration)-out cross validation, i.e., without six spectra of the two independent sample replicates at a time during the iterative validation process.

The regression was performed on the previously smoothed (Savitzky-Golay filter, 2nd order polynomial filter, 21 points) and multiplicative scatter corrected (MSC) spectra in the spectral range of 1,300–1,600 nm. The precision and accuracy of the developed PLSR model were evaluated by the coefficient of determination ($R^2$) and root mean square error (RMSE) of cross-validation.

Raw spectra, difference spectra, loading vectors of PCA analysis, and regression vector of PLSR analysis were examined in order to find and assign characteristic water absorbance bands showing considerable changes in response to changes in KCl concentration. Thus, identified bands were used to describe water spectral pattern of salt solutions. To visually represent changes of water spectral pattern as a function of salt concentration, different types of aquagrams were constructed, namely classic aquagrams, aquagrams with confidence intervals and temperature-based aquagrams. The instructions for all necessary calculations and steps to produce these charts are explained in a separate section (Water spectral pattern represented by aquagrams).

All data analysis was performed by using R Project for Statistical Computing (R Core Team, 2017) (RRID:SCR_001905) and an "aquap2" package (Pollner and Kovacs, 2016).

### Aquap2 Package

The "aquap2" package developed by Pollner and Kovacs (2016) (free download and instructions available at www.aquaphotomics.com) provides an easy-to-use data preparation and analysis tools developed for extending the functionalities of the R project software to the needs of aquaphotomics. It is a non-commercial, free-to-use software, which can dramatically speed up analysis time, especially in the case of large datasets. It is very flexible and allows an automation of highly repetitive tasks, while also providing special functionalities not available in other commercially available chemometrics software, such as frequently used graph—aquagrams.

Aquap2 package offers the following functionalities:

- Experimental design with randomization of samples, planned number of replicates, consecutives, and environmental control samples
- Data import from various file formats suited for a variety of spectral acquisition softwares
- Fusion of spectral data with data from data loggers monitoring the environment or sample holders
- Flexible data analysis customized for different grouping / splitting / slicing of data with encapsulated, i.e., stable color-coding of samples/groups
- Very flexible data visualization from raw spectra to automatically detected and labeled peaks in various multivariate models' outputs
- A variety of data pre-treatments (e.g., smoothing, standard normal variate transformation (SNV), multiplicative scatter correction (MSC), extended multiplicative scatter correction (EMSC), detrend transformation, derivatives (using different methods), averaging, resampling, artificial noise loading
- Chemometrics methods: principal component analysis (PCA), partial least squares regression (PLSR), soft independent modeling of class analogies (SIMCA) and different versions of aquagrams
- Different cross-validation and independent prediction options to support model optimization

## THE POWER OF RAW SPECTRA AND CONVENTIONAL SPECTROSCOPIC ANALYSIS

With so many chemometrics methods available, one often neglects the possibility that something can be extracted from the raw spectra, especially since changes in the water spectra in the near infrared region are subtle and difficult to observe with the naked eyes. However, the first, most natural step in all data analysis is to inspect the raw data.

In the NIR region, the water spectrum consists of four main maxima located approximately at 970, 1,190, 1,450, and 1,940 nm, which are due to the second overtone of the OH stretching band ($3\nu_{1,3}$), combination of the first overtone of the OH stretching and OH bending band ($2\nu_{1,3} + \nu_2$), the first overtone of the OH stretching band ($2\nu_{1,3}$) and combination of the OH stretching and OH bending band ($2\nu_{1,3} + \nu_2$), respectively (Luck, 1974). All these regions are informationally valuable. So far, more than 500 water absorbance bands have been identified under these broad peaks (Tsenkova, 2009; Tsenkova et al., 2015). Depending on the type of aqueous system, some regions can prove to be more suitable for analysis and provide more information; hence it is always advisable to closely examine each of these regions.

Let us now look at the raw, untreated spectra acquired for our potassium chloride example dataset (**Figure 2**).

The raw spectra were plotted to visualize the spectral changes introduced by adding different concentrations of salt to pure water. Two large peaks (around 1,450 and 1,940 nm attributed to the first overtone and combination region of OH stretching and bending vibrations) dominate the spectra of potassium chloride solutions. It is logical because salts do not exhibit the NIR spectra. Very small, broad features can also be observed around 1,190 nm. The region of the combination band shows significant noise due to the high absorption of water, which far exceeds 3 absorbance units and will be excluded from subsequent analysis. Further analysis will be performed only in the region of the first overtone of water, where for the most part, water absorbance bands can be clearly resolved and for which good literature sources exist about the specific assignments of water molecular conformations.

In this stage of data evaluation, two types of calculations are usually performed: averaging and spectral subtraction. The

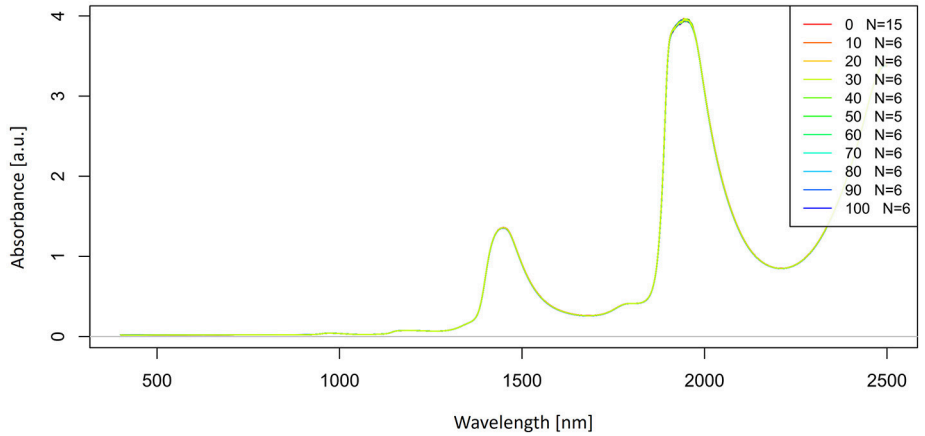



**FIGURE 2 |** Raw absorbance (logT-1) spectra in the entire spectral range of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM.

averaging can be done across all spectral consecutives and sample replicates. At this stage, the goal of averaging is to eliminate the influence of variations, which are not of primary interest, such as those attributable to different temperatures, humidity, or consecutive illumination. The average spectra of different groups of samples calculated this way will better reveal differences among different sample groups. However, the averaged spectra are influenced by outliers, so some measures of detecting and eliminating them should be taken before this step.

The next step is a spectral subtraction, which produces difference spectra. This is a very effective way for detection of subtle differences between the two spectra (Ozaki et al., 2003).

There are many approaches to spectral subtraction, and the simplest, classical approach is to subtract from the average spectrum of all samples, the averaged spectrum of pure water measured as a control during the experiment or of the solvent. This is the most simple and efficient method of bringing immediately a better visualization and observation of the water bands hidden under broad overtone and combination peaks.

Another subtraction method, recently developed, proposes a "closest spectrum" subtraction (Kojić et al., 2017). This subtraction method involves creating all the possible pairs of differences (solution—pure solvent) and finding the closest spectral pair (minimal difference) based on the smallest area under the curve of the difference spectrum. Thus, the found spectrum, the "closest spectrum," is then subtracted from the remaining spectra. Pure solvent spectra can be acquired during the experiment or found in a library of solvent spectra which must be previously created by performing an acquisition under various, mainly temperature, perturbations. This method provides, on average, a 4-fold increase in precision as compared to traditionally used average spectrum subtraction (Kojić et al., 2017).

Another way of enhancing differences is to calculate the difference spectrum along some perturbation of interest. This type of subtraction can reveal water absorbance bands activated by a particular perturbation. This simple approach, for example, allowed an immediate identification of main differences in the water structure between the groups of bacterial cultures *S. auerus* and *E. coli* (Nakakimura et al., 2012). In addition, in the study of the effect of soybean mosaic virus, the difference spectrum between the average spectra of healthy and diseased plants clearly revealed water absorbance bands due to virus-induced changes (Jinendra et al., 2010). Another example can be found in a study of the spectral behavior of mushrooms subjected to physical perturbation by different levels of mechanical vibration (Gowen et al., 2009b). The difference spectra obtained by subtracting the averaged spectrum of undamaged mushrooms from averaged spectra of damaged mushrooms subjected to different perturbation levels revealed sharp features around 1,398 nm for the two highest level of perturbations, which corresponds to absorption of free single water molecules trapped by ions (Kojić et al., 2014) at the mushroom surface originated from physically damaged cell walls.

Another highly efficient approach in revealing different water dynamics in samples is a subtraction of the 1st consecutive spectra from all other consecutive measurements. This subtraction technique was first applied in a study of different prion protein isoforms in water solutions (Tsenkova et al., 2004; Tsenkova, 2005), when it was shown for the first time that illumination changes the water system and each consecutive spectrum of the sample is influenced by light absorption. The effect of absorbed photons on water molecular systems increased a number of free water molecules available to interact with solutes in the aqueous system, performing "scanning" of solutes and the rest of the water molecular system resulting in changes of the corresponding spectra. In this way, additional information can be extracted, which is especially beneficial when the aqueous systems analyzed are very similar. In the case of the prion protein study, this approach revealed drastic differences in the free O-H absorbance bands and superoxides for different prion protein isoforms (Tsenkova et al., 2004; Tsenkova, 2005).

The spectra transformed as just described can also be further analyzed by using other data-mining approaches.

## SPECTRAL PREPROCESSING—IMPROVING AND ENHANCING SPECTRAL INFORMATION

The fundamental problem, not only in aquaphotomics analysis but also generally in all spectral analysis, is how to extract the useful information hidden in the complex spectral measurements. The objective of preprocessing is to enhance the information of interest, and decrease or remove unwanted influences on spectral signals.

The spectral preprocessing methods include mathematical pretreatments, such as centering and normalization (mean-centering, standard normal variate transformation (SNV)(Barnes et al., 1993); noise-reduction methods, such as smoothing or wavelet transform (Patil, 2015); baseline correction methods which include de-trending (Barnes et al., 1989); multiplicative scatter correction (MSC) (Dhanoa et al., 1994); extended multiplicative scatter correction (EMSC) (Martens and Martens, 2001); and spectral derivatives which, in addition to baseline correction, also resolve overlapping peaks.

Spectral patterns collected are usually affected by noise or instrumental variations that may have a detrimental effect on further analysis and conclusions that may be drawn (Gowen and Amigo, 2012). The weakly absorbing bands in the NIR region are far more affected as compared to the stronger ones. The best approach in ensuring high-quality and noiseless spectra, begins with the conditions of spectral collection which should be carefully controlled. Usually, collecting and averaging multiple scans successfully reduce the noise. However, some level of noise should be expected so that the common practice is to use smoothing techniques (Manley, 2014).

The most common de-noising techniques used in aquaphotomics methods are based on the Savitzky–Golay approach (Savitzky and Golay, 1964), which fits the spectral pattern to a polynomial function (second-order polynomial) in a step-wise manner. Continuous wavelet transform (CWT) is also one of the de-noising techniques, proved to be very efficient for processing analytical signals (Shao et al., 2003), and is of recently frequently used for enhancing spectral resolution and background removal in aquaphotomics works (Shao et al., 2010; Kang et al., 2011; Shan et al., 2015; Cui et al., 2016).

Mean centering of spectra is a pre-processing technique mostly used with principal component analysis (Agelet and Hurburgh Jr, 2010). It involves a subtraction of the average spectrum from the entire dataset, which results in reduced number of variables and complexity of subsequently built models (Manley, 2014).

Apart from random noises, the spectra of aqueous systems often exhibit baseline variations (in slope and offset) due to the scattering originated from differences in sample surface or particle size variations (Ozaki et al., 2003). Baseline offset problems are commonly solved by the application of SNV or MSC corrections methods. MSC is a better choice for correction when variations in the spectral slope are also present as a result of additive variation, which increases with wavelength due to the scattering present in samples. The disadvantage of MSC transformation is that it is sample-dependent; hence any change in the sample set requires a recalculation of all MSC related subsequent calculations (Dhanoa et al., 1994).

Detrending is also a possible choice for correction of baseline shift and curvilinearity. This method consists of modeling the baseline as a function of wavelength with a second-degree polynomial and a subsequent subtraction of this function from each spectrum individually.

With correction for baseline variations, one should be careful as sometimes they can contain information of interest. For example, in a study of prion protein isoforms, the benefit of multiplicative scatter correction was 2-fold. First, it confirmed the presence of scattering for one isoform of prion protein, which helped better understanding of its interaction with water by explaining that an increase in bulk water and changes in protein structure are the cause of scattering. Second, when correction for the scattering was applied, a subsequent analysis revealed differences in different protein isoforms not related to the scatter (Tsenkova et al., 2004). However, in a problem of somatic cell count determination, removal of the baseline variation by application of the second derivative transformation led to a diminished accuracy of prediction of somatic cell count in milk, leading to the conclusion that the baseline correction removed significant information (Tsenkova et al., 2001a).

The use of derivation as a pre-processing technique for NIR data is quite common. There are two ways of calculating derivatives: the Norris–Williams derivation (Norris and Williams, 1984) and Savitzky–Golay derivation (Savitzky and Golay, 1964). Derivatives can solve two basic problems with NIR spectra of aqueous systems: overlapping peaks and large baseline variations. The effect of derivatives is most clearly seen in the second derivative of a spectrum, which is able to separate overlapping bands. The second effect of the second derivative is removal of baseline shifts (Williams and Norris, 1987; Heise and Winzen, 2002). Two side effects of the derivatives are the loss of the original shape of a spectral curve, which may result in a difficult data interpretation and a reduction in signal-to-noise ratio. Choosing window size when performing derivatives should also be done with caution in the case of spectra of aqueous systems because this parameter influences a number of points in the resulting spectral vector (Rinnan et al., 2009), which may lead to a wavelength loss and a subsequent loss of information about some water bands.

Iwamoto et al. (1987) showed that the derivative transformation of spectra was a useful method of separating multiple absorptions in broad spectral peaks of water and used it successfully to better understand the state of water in foodstuffs. In aquaphotomics applications, the second derivative is a very popular and efficient approach for discovering activated water absorbance bands that are not visible in the original spectrum (see for example Jinendra et al., 2010; Jinendra, 2011; Kinoshita et al., 2012; Bázár et al., 2016; Kovacs et al., 2016).

Let us now look at the examples of application of these preprocessing steps on the spectra of potassium chloride solutions. The smoothed spectra were calculated by using a Savitzky-Golay filter (2nd order polynomial fit and 21 points) and presented in **Figure 3**. Only the area of the first overtone 1,300–1,600 nm is plotted to provide a better visualization of

how smooth the spectra should look. Next, a subtraction of the average spectrum of Milli-Q water from all the averaged spectra of potassium-chloride solutions was done and is presented in **Figure 4**.

The subtracted spectra reveled the existence of at least two major peaks under the broad overtone spectral curve of potassium-chloride solutions around 1,412 and 1,500 nm. It is also possible to observe a slight peak shift at 1,412 nm with increasing salt concentration.

The 2nd derivative spectra of potassium-chloride solutions were calculated by using a Savitzky-Golay filter (2nd order polynomial and 21 points) and presented in **Figure 5**. The second derivative spectra also indicate an existence of the band at 1,412 nm and we can also see the second band located at 1,462 nm.

With these simple preprocessing steps, we have so far identified at least two water absorbance bands activated by salt perturbation.

## CHEMOMETRICS- THE IMPORTANCE OF CONSISTENCY

Similar to the classical spectroscopy, the use of chemometrics methods is a crucial part of the aquaphotomics data analysis as well. It includes many well-known exploratory, classification and regression methods depending on the objective of the experiment.

Principal components analysis (PCA) (Cowe and McNicol, 1985) is one of the most useful and probably mostly commonly used exploratory technique in spectroscopy during the early stages of data analysis. Its objective is to determine a possible relationship between samples, i.e., to provide the first clues about major directions and sources of variation in the dataset. It compresses data by constructing new variables and the results are presented in scores and loadings plots. The scores plots visualize the spectra in the form of scores in the transformed space of newly constructed variables—principal components, while the corresponding loadings plots denote the contributions of original variables—wavelengths. The novelty of PCA application in aquaphotomics analysis is that a particular attention is given to the analysis of all loading vectors as they can reveal activated water absorbance bands.

PCA in the case of our salt dataset was used to describe multidimensional patterns in the spectral data and discover outliers. PCA data presented in the scores (**Figures 6**, **7**) and loadings plots (**Figure 8**) reveal major sources of variation in the data. The first two principal components describe more than 99.9% of variation in the dataset. The first principal component, whose loading shows two dominant features (a peak positive peak at 1,415 nm and a negative peak at 1,498 nm), is related to changes in water matrix caused by consecutive illumination. This effect is similar to that of temperature (Segtnan et al., 2001) in that free or weakly hydrogen bonded species absorbing at 1,415 nm increase at the expense of strongly hydrogen bonded water molecules absorbing at 1498 nm. The second principal component, which explains 11.403% of variation, shows the influence of concentration. It can be seen from the PC1-PC2 scores plot that while the scores move toward the negative part of the PC2 with increasing concentration, the pure water scores are entirely located in the positive part of this PC. The loading vector of PC2 presented in **Figure 8** reveals major water absorbance bands affected by the presence of salt in water i.e., 1,402, 1,444, and 1,530 nm. Regarding loading vectors, it is very important to look at all PC loadings since changes in water are very subtle and might be also described by a higher number of PC loading vectors.

The next steps of the analysis depend on the objective of the experiment. They can involve classification methods to group samples together according to their spectra, or regression methods to link sample spectra to some quantifiable properties (Roggo et al., 2007).The application of these methods in aquaphotomics analysis does not differ much as compared to the classical NIR applications. However, the unique characteristics for the aquaphotomics approach are as follows.

First, the initial step of the aquaphotomics approach involves qualitative analysis. This step may include the application of PCA or some unsupervised classification analysis, performed with the objective of data exploration and better understanding of spectral variability. This step may even include some preliminary regression analysis, which can show very poor prediction results and non-linearity existence. However, it can provide information about the existence of natural clusters of samples indicating the need for separate modeling for different groups of samples thus discovered. For example, the most accurate prediction of milk components such as protein, lactose and fat in cow milk was achieved when the models were separately built by using milk spectra from healthy and mastitis animals (Tsenkova et al., 2001a,c). A subtraction of the averaged spectra of these two groups will give us the first information about the "important" WAMACS to be used in further analysis. The presence of mastitis disease (bacterial infection) significantly alters the structure of water in milk and milk composition, causing non-linearity in the regression models if the spectra of healthy and mastitis animals are used together. In this case, separately built regression models form a part of the aquaphotome database, where a different regression model is applicable depending on the physiological status of the animal. In this respect, aquaphotomics does not aim nor considers it possible to build global models. This is especially true in the analysis of biological systems that are far too complex to be described with only one model.

Second, the most important feature of aquaphotomics analysis is the special attention paid to original and transformed spectral vectors as well as model outputs. This reveals the contribution of original variables—wavelengths, to model development and tracks consistently repeating variables. The identified variables with high contribution, which constantly repeat through all the steps of aquaphotomics analysis, are the most informative ones. For aquaphotomics, these variables are the places in the spectra, where various water molecular conformations absorb. Their identification is crucial for better understanding of the aqueous system and response of its water matrix to the perturbation. In other words, the variables, which consistently appear in all aquaphotomics analysis (i.e., in

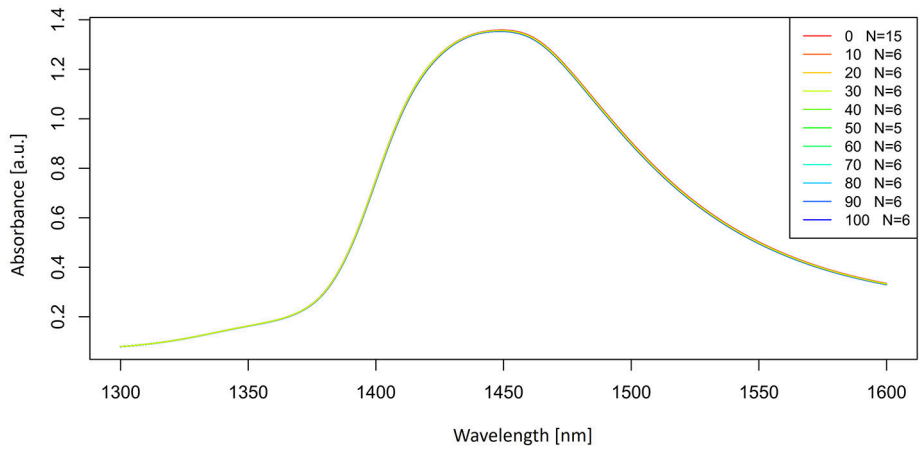



**FIGURE 3 |** Smoothed (calculated with a Savitzky-Golay filter using 21 points) absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM.

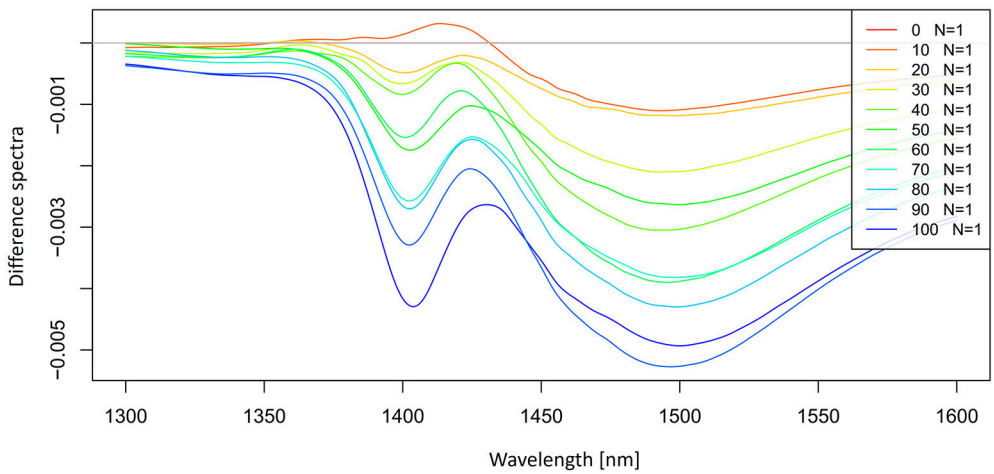



**FIGURE 4 |** Smoothed (calculated with a Savitzky-Golay filter using 21 points) average difference absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM. Average spectrum of Milli-Q water was subtracted from the spectra of potassium-chloride solutions.

subtracted spectra or transformed spectra, spectral derivatives, model outputs in the form of PCA loadings, PLSR regression vectors, SIMCA discriminating powers etc.), are the locations of water absorbance bands, where spectral variations under controlled and uncontrolled perturbations could be observed. If they persistently and consistently appear through all of the analysis, we can consider these water absorbance bands as activated.

Let us now look at the PLSR application on our salt dataset. The regression was performed on previously smoothed (Savitzky-Golay filter, 2nd order polynomial, 21 points) and MSC transformed spectra in the spectral range of 1,300–1,600 nm to build a model for prediction of potassium-chloride concentration. The results of PLSR analysis are presented in **Figures 9**, **10**, showing a close correlation and a relatively low error of cross-validation using five latent variables ($r^2$ = 0.9989, RMSECV = 1.147 mM, **Figure 9**). The main absorbance bands showing a significant weight in the PLS regression vector (**Figure 10**) match very well with those found in the previously applied methods, and all belong to the ranges of WAMACS found in the first overtone of water (Tsenkova, 2009). The favorable prediction results are not surprising since it is well established that salts influence the spectrum of water and these changes can be used for prediction of salt concentration (Grant et al., 1989; Gowen et al., 2015). Because salts do not absorb the NIR light, these results and the previously mentioned studies demonstrate the feasibility of aquaphotomics water-mirror approach. In other words, the absorbance bands of water can be used to obtain indirectly the information about changes in solute concentrations.

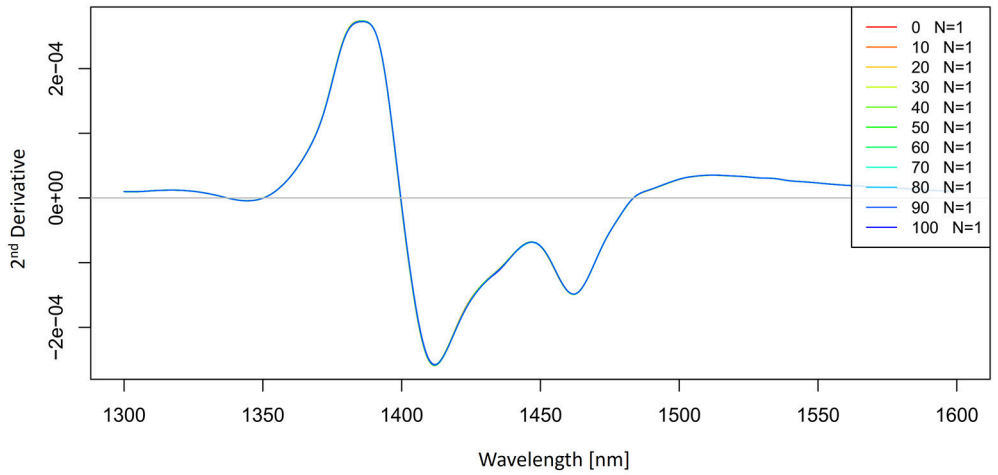



**FIGURE 5 |** 2nd derivative (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) average absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM.

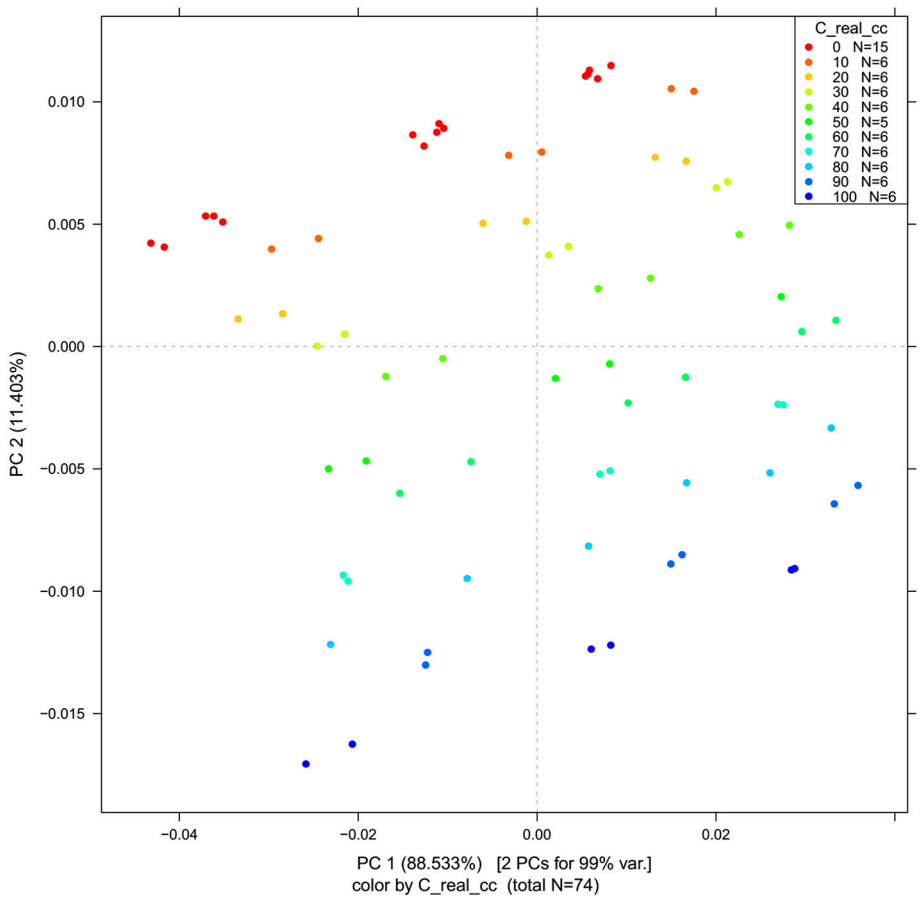



**FIGURE 6 |** PCA analysis of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM derived from the smoothed (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) and MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone)—Scores plots for the first two principal components.

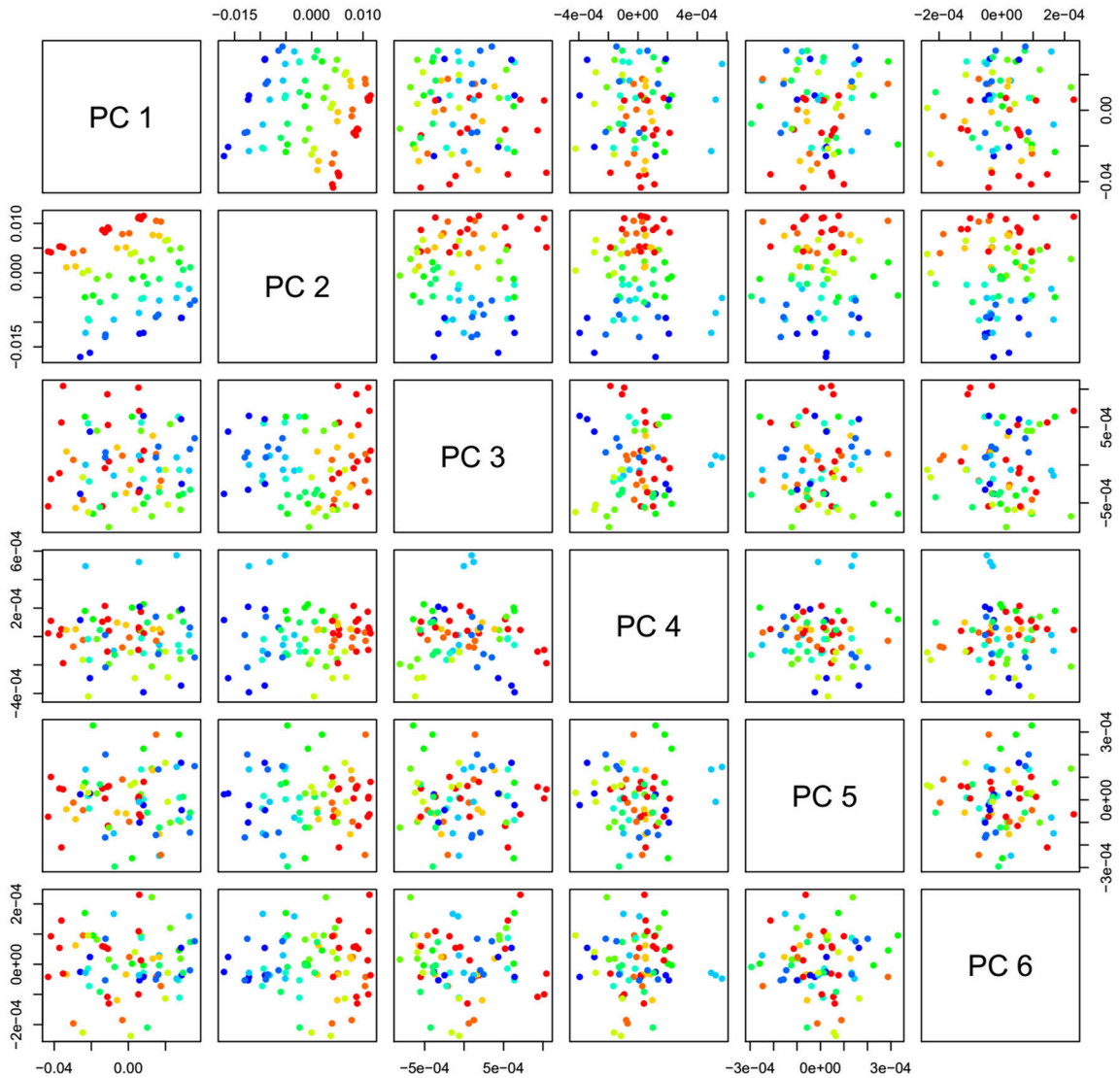



**FIGURE 7 |** PCA analysis of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM derived from the smoothed (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) and MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone)—Scores plots for the first six principal components.

It is worth mentioning that the analysis may include several more chemometrics methods that can also contribute to the identification of water absorbance bands activated by the perturbation of interest.

Employing discriminant analysis such as Partial Least Squares Discriminant Analysis (PLS-DA) (Martens and Martens, 2001) for discriminating between solvent and solutions can help in gaining more insight about how the solutes affect the water matrix of the solvent. For example, this method was employed to discriminate between solvent and pesticide–containing solutions (Gowen et al., 2011). Examination of the regression vectors of PLS discriminant analysis provides an additional help in revealing water absorbance bands activated by the presence of solutes.

Similarly, Soft Modeling of Class Analogies (SIMCA) (Wold and Sjöström, 1977) can be employed for the same purpose. The discriminating power of SIMCA analysis, in that case, reveals water absorbance bands with the highest discriminating power which distinguishes between pure solvent and solutions. One such example can be found in an aquaphotomics study concerned with measurements of different saccharides at millimolar concentrations (Bázár et al., 2015). Sometimes, both discrimination methods (SIMCA and PLS-DA) are employed for the same purpose of discriminating the solvent from the solutions and the discovery of additional information about activated water absorbance bands by solutes. In a study concerned with the detection of UVC damaged DNA, both PLS-DA and SIMCA were applied to distinguish between non-irradiated and

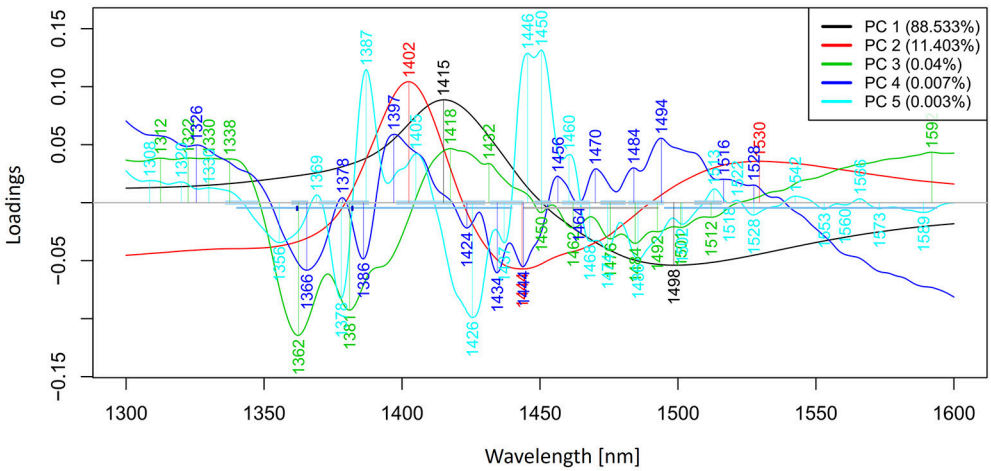



**FIGURE 8 |** PCA analysis of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM derived from the smoothed (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) and MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone)—Loadings plot.

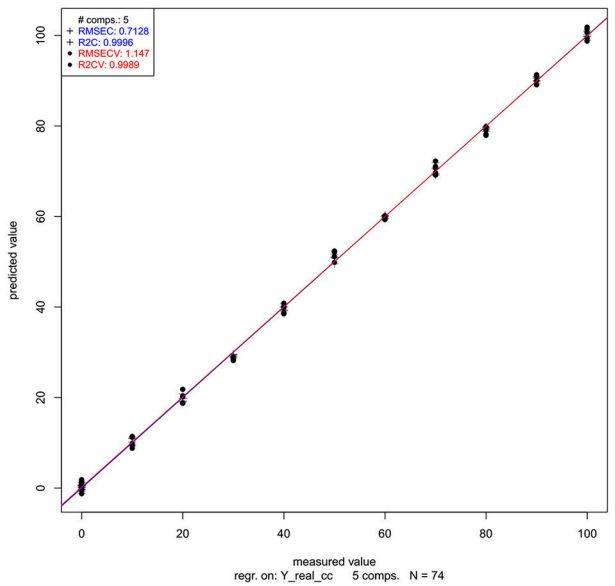



**FIGURE 9 |** PLSR analysis of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM derived from the smoothed (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) and MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) built for the prediction of potassium-chloride concentration: Y fit of training and one-sample-out cross-validation.

UVC-irradiated DNA solutions (Goto et al., 2015). Applying two chemometrics methods for the examination of one aspect of the experimental study demonstrates the stability of the applied methodology, namely, consistency in results.

Both the SIMCA and PLS-DA methods are naturally used in most cases when the objective of the study is discrimination between different samples. For classification and discrimination purposes in aquaphotomics, the most commonly used methods are SIMCA and PLS-DA. The SIMCA method was employed, e.g., for discrimination between healthy and mosaic virus infected soybean plants (Jinendra et al., 2010), for discrimination between healthy and mastitic animals based on the spectra of urine, blood and milk of dairy cows (Tsenkova, 2004), for discrimination between different brands of commercially available mineral waters (Munćan et al., 2014), for discrimination of different bacteria strains (Remagni et al., 2013; Slavchev et al., 2015, 2017) and others. The PLS-based discriminant analysis was applied for discrimination between irradiated and non-irradiated DNA solutions (Goto et al., 2015), discrimination between solvents and pesticides containing solutions (Gowen et al., 2011), and discrimination between worn and new soft contact lenses based on conventional hydrogels (Šakota Rosić et al., 2016).

Quantitative aquaphotomics analysis usually includes partial least squares regression (PLSR) (Martens and Martens, 2001) or principal component regression (PCR) (Næs et al., 2002). The principal uniqueness of the aquaphotomics approach in the application of these two methods is the utilization of water absorbance bands for indirect quantification of analytes in water, which change the water matrix. The feasibility of this approach was demonstrated in a study whose objective was quantification of different types of salt in water solutions (NaCl, KCl, $MgCl_2$, and $AlCl_3$), where the overall detection limit of 1,000 ppm was reported (Gowen et al., 2015). The experiment was reproduced in three independent laboratories by using 3 different spectrometer systems and in different ambient conditions. The reported detection limit of 1,000 ppm indicates that under specified conditions, the aquaphotomics approach substantially improved the detection limit for NIRS (around 5 times) (Pasquini, 2018).

Using an aquaphotomics approach, PLSR gave excellent results for quantification of various analytes in water solutions such as sugars [glucose, fructose, sucrose and lactose and their

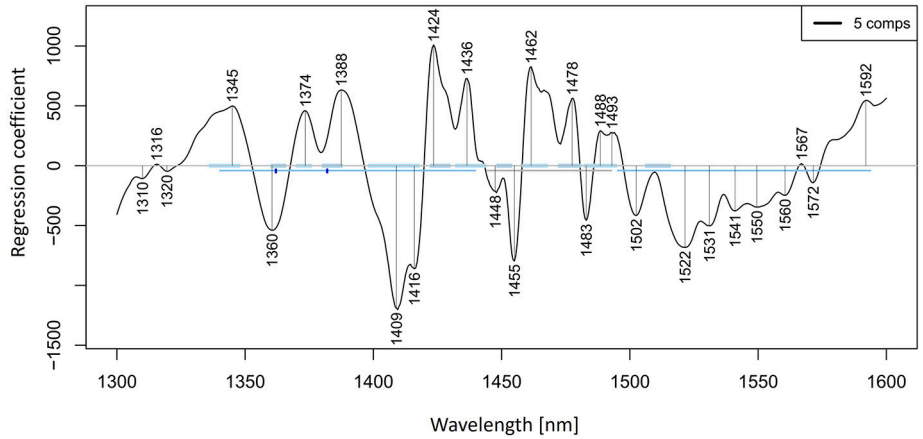



**FIGURE 10 |** PLSR analysis of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM derived from the smoothed (calculated with a Savitzky-Golay filter using 2nd order polynomial and 21 points) and MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) built for the prediction of potassium-chloride concentration: Regression vector.

mixtures (total sugar and each sugar concentrations)] (Bázár et al., 2015), insulin protein (Chatani et al., 2014), DNA, isolated cyclobutane pyrimidine dimers, and UVC-irradiation dose (Goto et al., 2015). The same approach also provided a favorable accuracy of measurements in more complex biological samples, such as human serum albumin (HSA) and γ-globulin in phosphate buffer solutions (Murayama et al., 1998), urinary estrone-3-glucuronide ($E_1G$) concentrations in urine of giant pandas (Kinoshita et al., 2010, 2012), HIV virus in human plasma (Sakudo et al., 2005), somatic cell counts in cow milk samples (Tsenkova et al., 2001a; Tsenkova, 2004), as well as fat, lactose, protein and urea nitrogen content of milk (Tsenkova, 2004).

Very recently, a critical review on NIRS and its modern perspectives expressed concerns regarding the capability of aquaphotomics for measurement of analytes in very low concentrations, given the fact that the concentrations of 5,000 ppm (mg $L^{-1}$) or 0.5% (w/v) are roughly regarded as a common limit of quantification for NIRS (Pasquini, 2018). Capability comparison of the traditional NIRS and aquaphotomics approach is based on an incorrectly assumed equivalence. While the established limit of detection for the traditional approach is based on the utilization of absorbance bands of analytes in the NIR region, the aquaphotomics approach utilizes water absorbance bands. In this sense, the quantification of analytes is based on entirely different principles, and as such, logically offers different limits of detection. Different approaches and their accuracy of detection were well demonstrated in studies on the measurement of concentrations of polystyrene particles in water (Tsenkova et al., 2007b). When the first overtone of water (i.e., aquaphotomics approach) was used to develop a model for low concentrations of polystyrene particles in aqueous suspensions (1 – 0.0001%), the measurements achieved a high accuracy even in the case of very low concentrations. However, when the traditional approach was applied and measurements were based on the polystyrene band near 1,680 nm (C-H stretching from aromatic C-H (2ν) (Workman, 2016)—i.e., decreasing particle concentration led to a substantial decrease in accuracy of prediction.

Aquaphotomics can work with very water-rich systems. The intensity of water bands in the NIR spectra of such systems is much stronger than that of any constituent (Tsenkova, 2004), especially if they are in very low concentrations. The possibility of detecting and measuring such low concentrations arises from the fact that every molecule of analyte is hydrated with an abundance of water molecules, which adapt to its structure and assume various conformations that can be observed based on their respective absorbance bands in the NIR region. Since many water molecules are involved with hydration of just one molecule of analyte, the water acts as a sort of amplifier, and instead of measuring analytes directly, the information on their concentration is obtained indirectly by measuring changes in always abundant solvent molecules.

NIR spectroscopy as a non-destructive tool offers the advantage of *in vivo* spectral monitoring of living objects. Aquaphotomics combined with time-resolved NIR spectroscopy allows a better understanding of biological functions and underlying water dynamics.

One of the excellent methods for exploring water dynamics is generalized two-dimensional (2D) correlation spectroscopy (Noda et al., 1995; Liu et al., 1996). In 2D correlation spectroscopy, an external perturbation is applied to a system during spectral measurements, which enables exploration of spectral signals as a function of time or perturbation level (where perturbation can be a number of consecutives, temperature, concentration etc.). This method has significant advantages over one-dimensional spectra. Spreading the spectral region over another dimension allows a deconvolution of overlapped bands and monitoring a specific order of spectral intensity changes. Moreover, 2D correlation spectroscopy offers the possibility of investigating various intra- and inter-molecular interactions through selective correlation of peaks. This technique, in addition to PCA, considerably contributed to the understanding of the

structure of liquid water (Segtnan et al., 2001). Furthermore, it was applied for extraction of useful information from NIR spectra of protein aqueous solutions during heat-induced denaturation of ovalbumin (Wang et al., 1998) and acid-induced denaturation of human serum albumin (Murayama et al., 2000). The method can be applied even in the case of complex biological fluids such as milk (Czarnik-Matusewicz et al., 1999; Tsenkova, 2004) or complex biological samples such as fruits (Giangiacomo et al., 2009). 2D correlation analysis was also employed for the investigation of wafer etchant solutions composed of several inorganic acids (HCl, $H_2SO_4$, $H_3PO_4$, and $HNO_3$) (Chang et al., 2018). This study, using a typical water-mirror approach, applied 2D correlation analysis to examine NIR water bands perturbed by four acids and determined their dissimilar characteristics. The results showed that components with higher acidity in single-component samples perturbed water hydrogen bond network more significantly, and in turn allowed more accurate concentration measurements. Heterospectral correlation (Noda and Ozaki, 2004) i.e., investigation of correlation between water absorbance bands in different regions of the electromagnetic spectrum (IR and NIR) or by different techniques (NIR and Raman spectroscopy) can significantly contribute to the development of aquaphotomics through discovery and identification of new water absorbance bands. However, it should be pointed out that there is one inherent weakness of the method, i.e., high level of sensitivity to noise.

Other approaches for examination of water dynamics are also often in use. For example, plotting SIMCA interclass distance as a function of time revealed time-dependent spectral dynamics of virus infection in soybean plants (Jinendra et al., 2010). The SIMCA interclass distance between the groups of infected and non-infected plants showed small values of around 1.2 (2 weeks after inoculation), then gradually decreased to the lowest value of 0.8 (3 weeks after inoculation). After this critical point, the value of interclass distance increased steadily. Thus, revealed water dynamics mirrored the dynamics of viral infection where, due to the defense reaction from the plants, the disease impact was initially suppressed exactly 3 weeks after inoculation. The same approach was utilized in a study of the ovulation period in giant pandas (Kinoshita et al., 2010). Interclass distances were calculated between spectra of urine collected each day in the time series and urine spectra collected at the first day of investigation when the female animals had been in an estrous state. This analysis showed that the SIMCA distance between these two groups increased simultaneously with an increase in E1G concentration, a major estrogen metabolite excreted in the urine during estrus. Another study was concerned with investigation of protein fibrillation and employed spectral monitoring of water structural changes in real time during fibrillation of insulin (Chatani et al., 2014). This study monitored the process of fibrillation of insulin indirectly by monitoring water molecular structure dynamics in the region of the first overtone (1,300–1,600 nm), while the verification of formation of fibrils was performed by two methods i.e., FTIR spectroscopy and Atomic Force Microscopy. The PCA analysis of NIR spectra of protein solutions found that for the first two PCs, score changes can be mainly attributed to a change in light scattering; however, the scores of PC3, when expressed as a function of time (in minutes), showed a time course of changes in water structure coinciding well with the proposed nucleation, elongation and equilibrium phase of protein fibrillation (Chatani et al., 2014). It is worth mentioning that other ways of exploring water dynamics are possible. For example, expressing SIMCA interclass distance as a function of consecutive illumination or temperature can reveal different responses to perturbation in different samples, which otherwise, without perturbation, may be difficult to discriminate due to a high similarity. Also, expressing SIMCA interclass distance between solvent and solutions of varying concentrations, as a function of concentration, may reveal concentration ranges in which solutes have structure-breaking and structure-making effect, thus indicating the need for building separate regression models for different ranges of concentrations.

Recently, several novel chemometrics methods were introduced to aquaphotomics studies. Multivariate curve resolution-alternating least squares (MCR-ALS) was applied to characterize the effects of temperature and salt perturbations on the NIR spectra of water in order to gain more insight into hydrogen bonding (Gowen et al., 2013). This advanced data analysis technique applies a factor model approach with the objective of recovering pure concentration and spectral profiles of the components in complex mixture systems without any prior knowledge of these features (Czarnecki et al., 2015). To perform MCR, however, one has to estimate firstly a number of significant components, usually based on PCA analysis, In contrast to PCA, MCR can provide results that have actual physical and chemical meaning (Czarnecki et al., 2015). The "components" in terms of water structures could be interpreted as the changing forms of water when perturbations were applied. Three distinct components were found with varying temperature dependence in the range 30-45°C in the region of first overtone of water, while different salts and salt concentration levels affected the water hydrogen bonded network in different ways according to its acidity (Gowen et al., 2013). By resolving different systems into idealized pure components, MCR-ALS allowed better examination of water molecular matrix and resulted in the conclusion that the water structure can be reasonably interpreted as a multi-state system.

Evolving factor analysis (EFA) was applied for exploration of hydration and secondary structures of bovine serum albumin in aqueous solutions (Yuan et al., 2003). Application of this method allowed an extraction of spectral information, which indicated significant changes of bovine serum albumin in secondary structure. The application of independent component analysis (ICA) was reported in spectroscopic analysis of hydrogen bonding in water-acetone mixtures for resolving the spectra to independent components and obtaining their concentration profiles (Monakhova et al., 2014). A Gaussian fitting method was applied to study glucose-induced variation of water under temperature perturbation (Cui et al., 2016). This method, applied on a NIR difference absorbance spectra (in region 700–1,100 nm), helped identify and quantify 16 inorganic salts in water in the concentration range from 30 to 500 mM (Steen et al., 2015).

A series of articles were also published on employing and developing various chemometrics methods specifically for

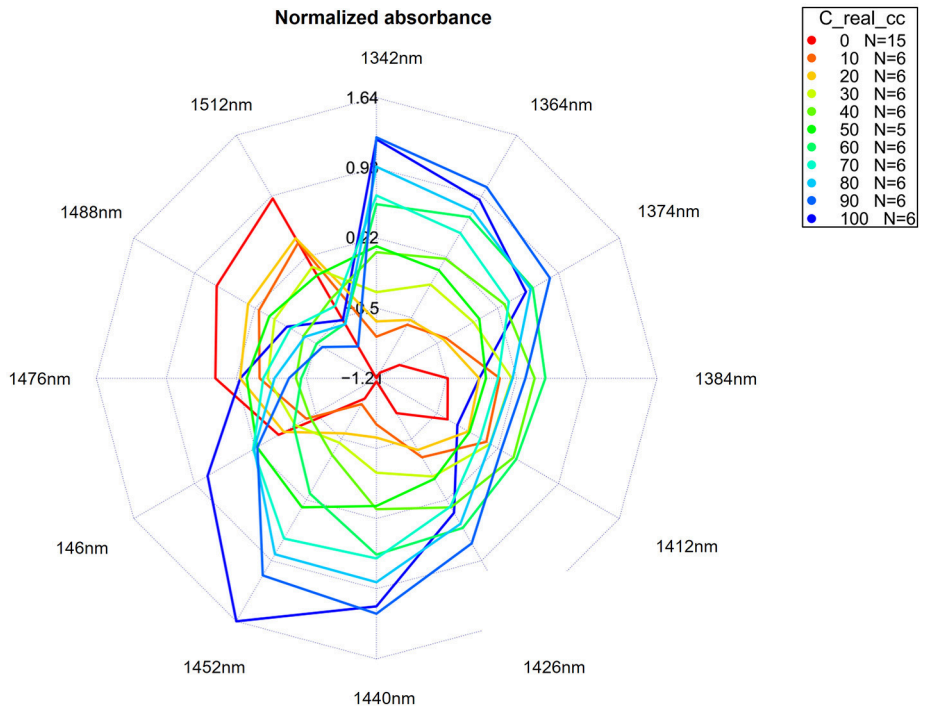



**FIGURE 11 |** Aquagrams without confidence intervals of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM calculated on the MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) using the "classic" mode.

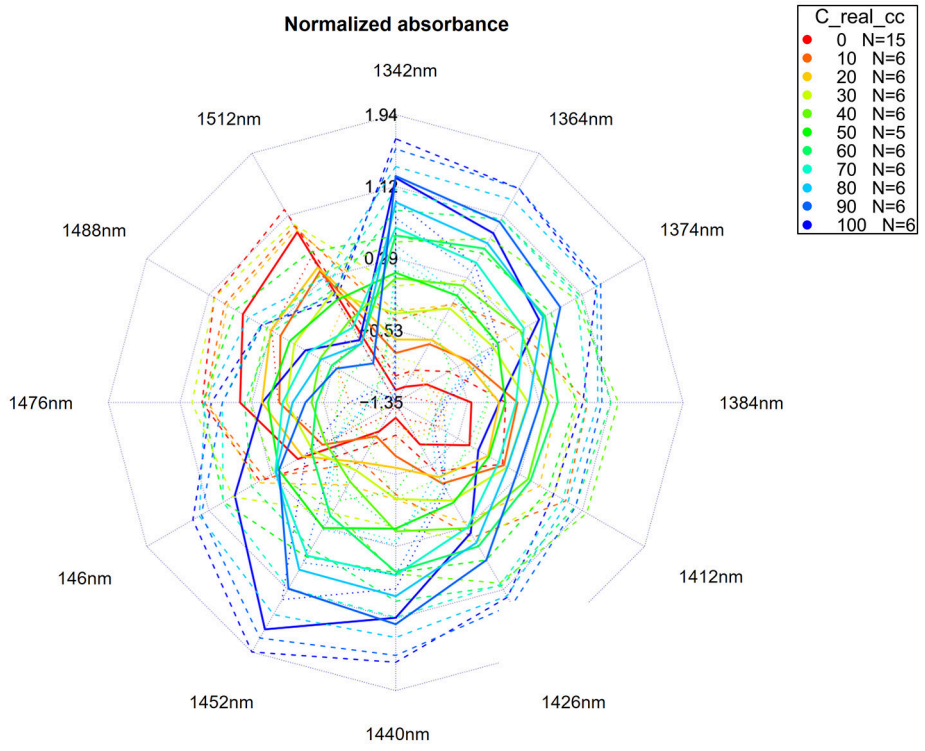



**FIGURE 12 |** Aquagrams with 95% confidence intervals of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM calculated on the MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) using the "classic" mode.

temperature-perturbed samples (Peinado et al., 2006; Shao et al., 2010, 2018; Kang et al., 2011; Shan et al., 2015; Cui et al., 2017b). Instead of trying to eliminate the influence of temperature, a Parallel Factor (PARAFAC) model was used to extract and separate relevant sources of both physical and chemical information (Peinado et al., 2006). PARAFAC analysis was also used to rationalize concentration-dependent peak shifts and quantification of different water species in acetone (Andrews et al., 2014), and also for a quantitative analysis of the NIR spectra of temperature-perturbed mixtures, water-ethanol-propanol and water-ethanol-glycerin (Peinado et al., 2006). Multilevel simultaneous component analysis (MSCA) has been applied to the investigation of a relationship between temperature and NIR spectra of different samples in different concentrations: water-ethanol-isopropanol, (Shan et al., 2015) and water-glucose (Cui et al., 2017a) under temperature-perturbation. This method was proposed specifically for analyzing multivariate data at different levels (Timmerman, 2006). The method offers a unique way to study the composition of solvent, temperature effect and quantitative analysis (Shan et al., 2015). Cui et al. tested three high-order chemometric algorithms: multiway principal component analysis (MPCA) (Wold et al., 1987), parallel factor analysis (PARAFAC) (Bro, 1997) and alternating trilinear decomposition (ATLD) (Wu et al., 1998) in the analysis of temperature-dependent NIR spectra of binary and ternary water-alcohol mixtures (Cui et al., 2017b). All three algorithms proved to be very powerful tools for capturing temperature– and concentration–induced spectral variations, from which a structural variation could be observed and a quantitative determination performed. Another work of Shao et al. proposes mutual factor analysis (MFA) for quantification based on temperature-dependent NIR spectra (Shao et al., 2018). In this work, multi-component mixtures were analyzed for quantification of components and better understanding of molecular interactions in solutions. From the spectra of water–glucose mixtures, both spectral variations induced by temperature and concentration were obtained while serum samples were used for method validation (Shao et al., 2018).

The ultimate choice of chemometrics method to be applied in aquaphotomics analysis depends on the type of the aqueous system explored, spectral dataset and the research objective. Obviously, there are many chemometric methods available. The important aspect of every aquaphotomics analysis is emphasis on consistency so that each preprocessing method, conventional spectroscopic method or chemometrics method applied to extract the information from water spectra can contribute to the development of an emerging aquaphotome. Each step of aquaphotomics data analysis is important, because it can contribute to better understanding of the complexity of aqueous systems, irrespective of chemometrics method applied.

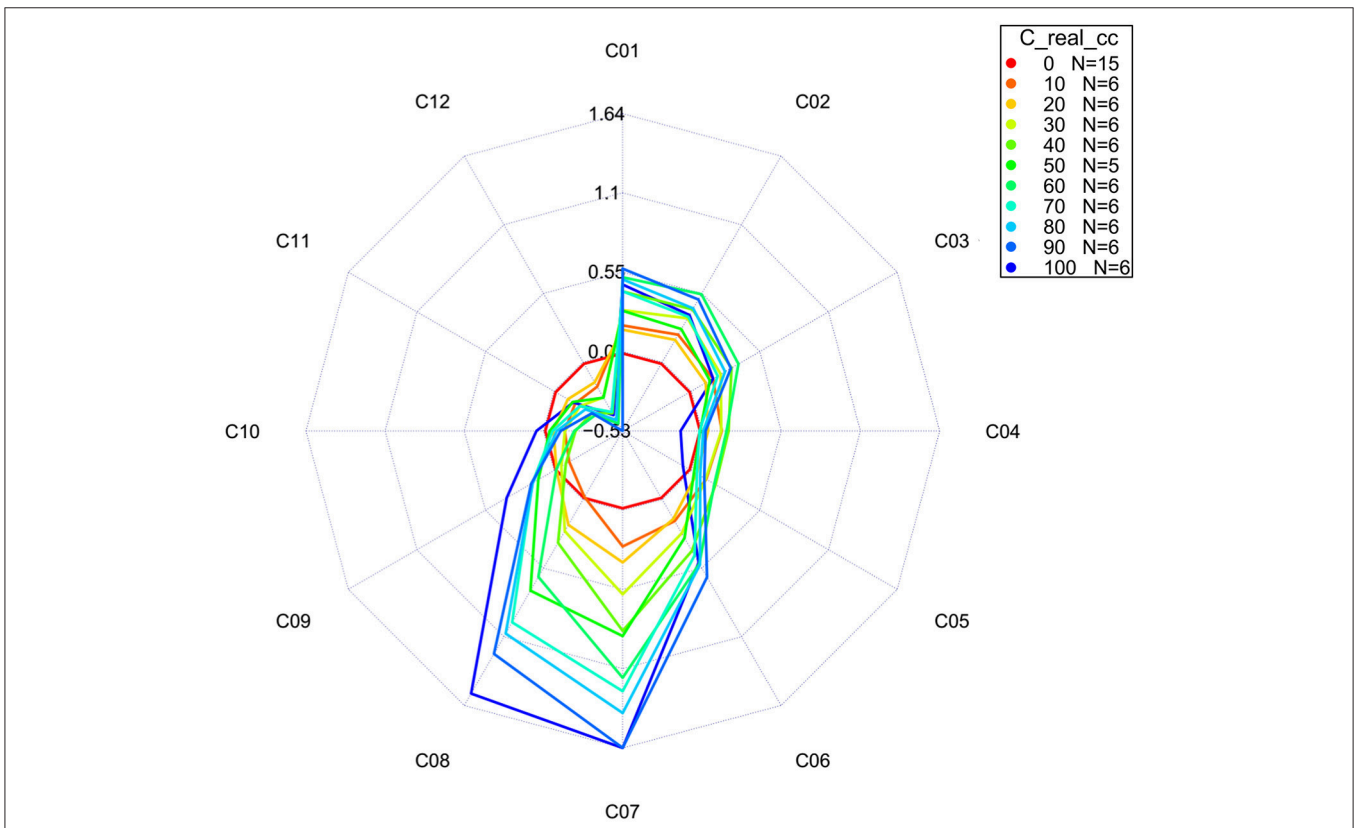



**FIGURE 13 |** Aquagrams without confidence intervals of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM calculated on the MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) using the "temperature-based" mode.

With reference to our example of potassium chloride solutions, after examining the raw spectra, difference spectra, second derivative spectra, loadings of PCA analysis and regression vector of PLSR analysis, we have identified the main water absorbance bands activated by the perturbation of potassium chloride in the concentrations up to 100 mM. The last step of analysis for our worked-out example is to represent water absorbance spectral patterns using aquagrams.

## WATER SPECTRAL PATTERN REPRESENTED BY AQUAGRAMS

### Classic Aquagrams

In data analysis, many situations arise where data visualization is helpful, even essential, for better understanding. In aquaphotomics, the need arose for a clear and comprehensive graphical representation of the water spectral patterns as well as for their easy comparison. That is why the aquagrams were introduced (Tsenkova, 2010).

When activated water absorbance bands are found based on the previously described steps, the last step is to apply MSC or SNV transformation of the raw spectra, and extract the absorbance at selected activated water bands. Thus, the calculated absorbance is normalized and averaged for different samples or sample groups, and the values are displayed on radial axes defined by the activated water absorbance bands in a radar chart.

The normalized absorbance is calculated as follows:

$$A'_{\lambda} = \frac{A_{\lambda} - \mu_{\lambda}}{\sigma_{\lambda}} \quad (1)$$

Where $A'_{\lambda}$ - is a normalized absorbance displayed on the aquagram, $A_{\lambda}$ - absorbance after multiplicative scatter correction (MSC) or standard normal variate transformation (SNV), $\mu_{\lambda}$ – mean of all spectra for the examined group of samples after transformation, $\sigma_{\lambda}$ – standard deviation of all spectra for the examined group of samples after transformation, $\lambda$ – selected wavelengths chosen for display from activated water absorbance bands.

An exact number of axes as well as water absorbance bands will be chosen for display, depending on a specific system and perturbation; however, the axes always display various conformations of water molecules, making aquagrams very convenient tools for a quick insight into the water structure of the system. For the first overtone of water, the axes of the aquagram are usually based on previously discovered 12 WAMACs. The aquagrams are visually very convenient to allow a fast and comprehensive comparison of different systems or conditions of the same system by comparison of their WASPs.

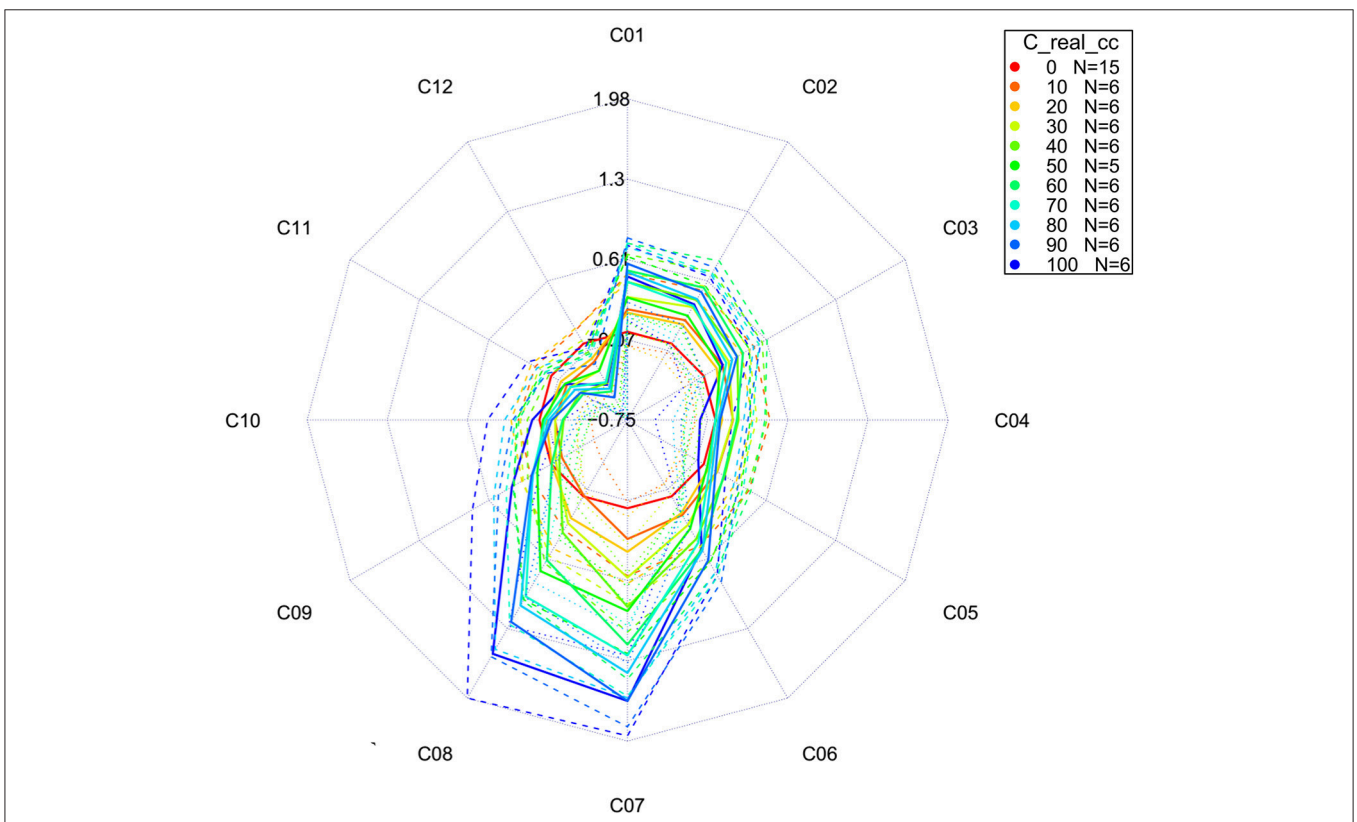



**FIGURE 14 |** Aquagrams with 95% confidence intervals of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM calculated on the MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) using the "temperature-based" mode.

As it can be seen from Equation (1), the classic aquagram is a relative construction, depending on the samples included in calculation. Also, it is a matter of choice whether the display of absorbance calculated based on the above equation is done by using a circular chart (radar chart) or a linear one. The package aquap2 offers both options (Pollner and Kovacs, 2016).

The more advanced version of a classic aquagram is an aquagram with confidence intervals (Pollner and Kovacs, 2016). This aquagram adds one more function, the possibility to observe whether the differences among WASPs presented in the aquagrams are statistically significant. This type of aquagram, in addition to averaged WASPs for selected groups of samples, displays its confidence intervals with 95% upper and lower limits, as calculated by using the Bootstrap method for data validation and uncertainty estimation (Davison and Hinkley, 1997; Pollner and Kovacs, 2016). With this novel function, the aquagrams with confidence intervals are not only convenient for visualization, but also especially suitable for classification and discrimination.

For our example dataset of potassium chloride solutions, after selecting wavelengths from the WAMACS regions in the 1st overtone of water based on the previous steps of the analysis, the classic aquagrams without and with confidence intervals, calculated by using aquap2 package, are presented in **Figures 11**, **12**.

In both types of aquagrams, it is easy to observe a large difference between the spectral patterns of water (red line) and salt solutions. Increasing the concentration of salt in water leads to increased absorbance in the region between 1,342 and 1,374 nm which corresponds to $C_1$, $C_2$, and $C_3$ WAMACS, i.e., absorbance of the free OH stretch (OH-(H2O)n, $n = 1 \ldots 4$) (Xantheas, 1995; Robertson et al., 2003). An increase in the absorbance with increasing salt concentration can also be seen in the region stretching from 1,440 to 1,452 nm, i.e., $C_7$-$C_8$ WAMACS that are known as bands of water hydration (Gowen et al., 2009a) and water dimers ($S_1$) (Segtnan et al., 2001; Cattaneo et al., 2009) and symmetric and asymmetric stretching of the first overtone of water (Siesler et al., 2008; Cattaneo et al., 2009; Gowen et al., 2009a). However, in the range between 1,476 and 1,512 nm, i.e., $C_{10}$-$C_{12}$, samples with higher salt concentration show lower absorbance values and this region is usually connected to strongly hydrogen bonded water (Segtnan et al., 2001; Tsenkova, 2009). The spectral pattern of salt solutions represented in the aquagrams shows that for the range of concentrations of salt under study, increasing salt concentration has a structure-breaking effect on water.

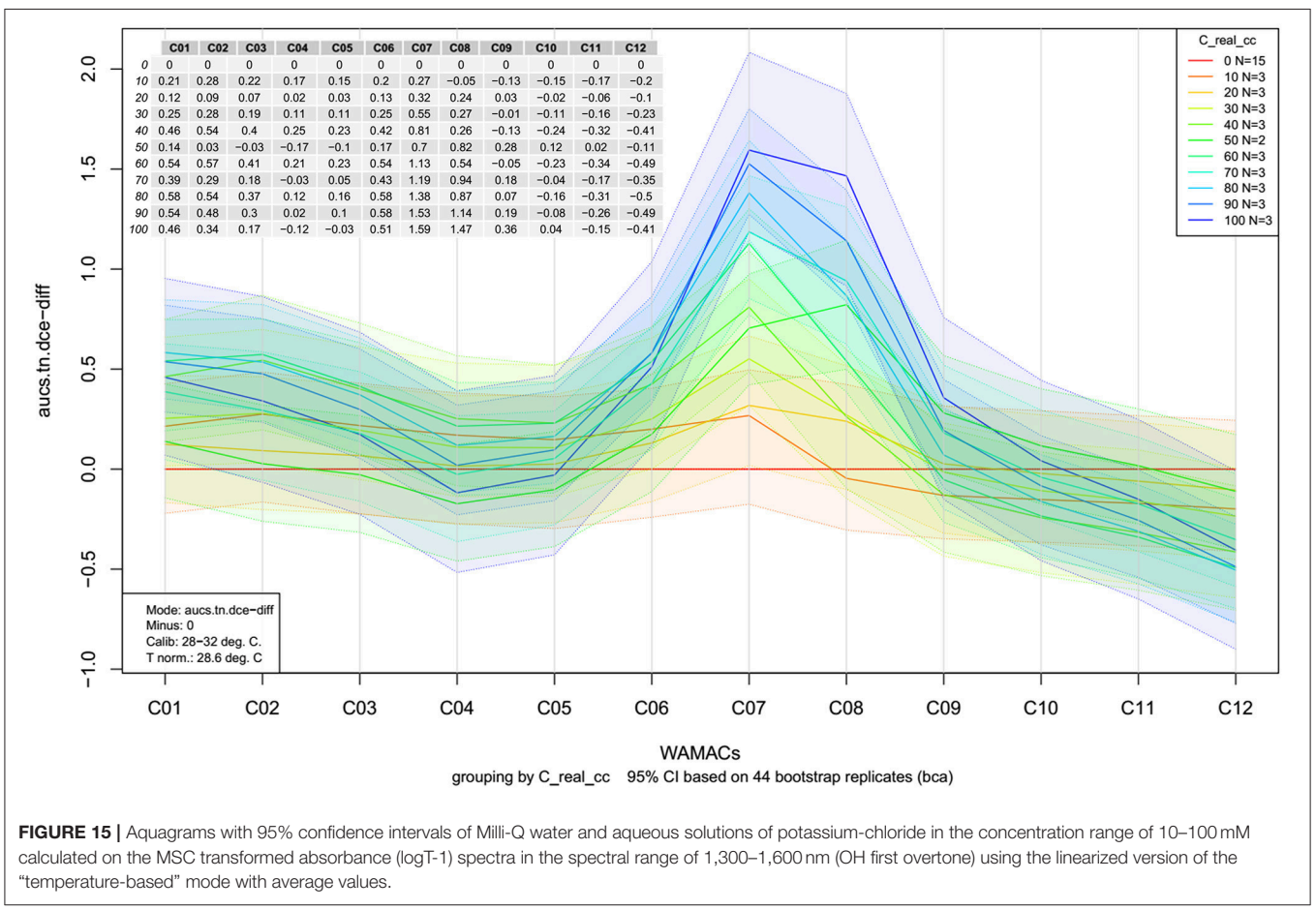



**FIGURE 15 |** Aquagrams with 95% confidence intervals of Milli-Q water and aqueous solutions of potassium-chloride in the concentration range of 10–100 mM calculated on the MSC transformed absorbance (logT-1) spectra in the spectral range of 1,300–1,600 nm (OH first overtone) using the linearized version of the "temperature-based" mode with average values.

## Temperature-Based Aquagrams

In the previous section, we briefly mentioned that classic aquagrams are relative constructs, meaning that the WASPs displayed depend on the samples included in calculation. This is disadvantageous if the WASPs of samples or groups of samples ought to be compared over time or in different experiments. The development of a new temperature-based aquagram (Pollner and Kovacs, 2016) overcomes this difficulty by transformation of how spectral changes are expressed.

For the calculation of temperature-based aquagrams, it is necessary to first acquire a spectral library consisting of spectra of pure water (Milli-Q) at different temperatures covering a wider range of temperatures than the one expected to be used during the experiment. This created library, or so-called *reference dataset*, provides the basis for temperature aquagram calculation. The spectra from this dataset are to be compared with the spectra acquired during the experiment—*experimental dataset,* giving the ground to express the effect of certain perturbation on spectral pattern of experimental samples in terms of the effect of temperature on pure water spectra. In this way, the effect of any perturbation on samples can be expressed in the "temperature equivalent units," in other words, changes in pure water spectra caused by temperature.

The calculation of a temperature-based aquagram is based on a comparison of areas covered by 12 WAMACS (Ci, $i = 1, 12$) coordinates in the region of the 1st overtone of water. The average spectra across all sample replicates and consecutive scans are calculated for the reference and experimental datasets. The area under the curve (AUC) for every single average spectrum for both reference and experimental datasets, at the wavelength range of each WAMACS (Ci) is calculated by taking into account the baseline estimated by linear fitting on the two edges of the first overtone region (i.e. through 1,300 and 1,600 nm points). The ratio of AUCs for every single water matrix coordinate and AUC for the first overtone region (i.e., 1,300–1,600 nm) are calculated for each averaged spectrum of both datasets in order to provide normalized values for comparison of reference and experimental datasets and to eliminate possible differences due to the scattering or path length differences. Using local polynomial regression for the reference dataset, a continuous array of values for the relative area of each Ci is calculated for a continuous temperature range chosen to include a specific temperature. In this way, a temperature calibration equation is obtained establishing a relationship between temperature and each Ci area, including the temperature at which the experiment was performed. When it is known how each Ci area for the pure water dataset is changed as a function of temperature, it is possible to pair these changes to spectral changes in the experimental dataset, i.e., to perform linking (mapping) and express the changes in Ci areas of the experimental datasets in the unit of temperature (degree Celsius) equivalent.

With this type of aquagram, it is also possible to include confidence interval limits. In that case, it is also necessary to perform transformation of upper and lower 95% confidence limits in the same manner just described above for the average spectra from the experimental dataset.

The whole calculation procedure for temperature-based aquagrams is implemented in the aquap2 package of R programing language (Pollner and Kovacs, 2016; R Core Team, 2017). An obvious disadvantage of temperature-based aquagrams is that they are based on previously discovered WAMACS regions in the first overtone of water (Tsenkova, 2010), meaning that at the moment this type of aquagram cannot be used for other windows of the electromagnetic spectrum where water absorbs.

The temperature based aquagrams without and with confidence intervals for our dataset of aqueous solutions of potassium-chloride spectra are presented in **Figures 13**, **14**, respectively. The linearized version of the temperature-based aquagram for **Figure 14** is plotted in **Figure 15**, where the additional table shows average values at all WAMACs.

Further understanding can be obtained from the temperature-based aquagram. The addition of, for instance, 90 mM potassium-chloride to Milli-Q water results in structural changes equivalent to temperature changes of about 0.54, 0.48, 0.3, 0.02, 0.1, 0.58, 1.53, 1.14, 0.19, −0.08, −0.26 and −0.49°C at $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $C_6$, $C_7$, $C_8$, $C_9$, $C_{10}$, $C_{11}$, and $C_{12}$ coordinates, respectively. Furthermore, the differences are statistically significant for calculated confidence intervals, e.g., the above listed differences between pure Milli-Q and 90 mM aqueous solution of potassium-chloride was significant ($p < 0.05$) at the coordinates $C_1$, $C_2$, $C_3$, $C_6$, $C_7$, $C_8$, $C_{11}$, and $C_{12}$.

## CONCLUDING REMARKS

In this paper, the fundamentals of the aquaphotomics approach to data analysis have been presented and discussed. A variety of applications illustrate the potential of aquaphotomics as a powerful new spectroscopic tool to study various aspects of aqueous and biological systems, which are of interest in the pharmaceutical and biomedical fields. The process of analysis illustrated by the application of aquaphotomics analysis on aqueous salt solutions was intended as guidance for certain steps of the analysis with the simplest experimental system, which anyone can easily reproduce. Together with the examples from sources of literature referenced throughout the text, this paper should provide the basis for independent experimental work in this field. The existing aquaphotomics literature shows the results which are probably only the tip of the iceberg of possible applications. With the explained methodology of aquaphotomics analysis presented herein, we hope that scientists and chemometricians will implement it in their fields and come up with new ideas of applications as well as new and more sophisticated mathematical tools to contribute to this growing field.

## AUTHOR CONTRIBUTIONS

ZK, BP, and RT designed and performed experiments. ZK performed data analysis. JM, ZK, and RT interpreted results and wrote the manuscript.

## ACKNOWLEDGMENTS

The author JM gratefully acknowledges the financial support of JSPS Postdoctoral Fellowship for Foreign Researchers (P17406).

The author ZK gratefully acknowledges the support by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences, Hungary and by the ÚNKP-17-4 New National Excellence Program of the Ministry of Human Capacities.

## REFERENCES

Agelet, L. E., and Hurburgh Jr, C. R. (2010). A tutorial on near infrared spectroscopy and its calibration. *Crit. Rev. Anal. Chem.* 40, 246–260. doi: 10.1080/10408347.2010.515468

Andrews, N. L., MacLean, A. G., Saunders, J. E., Barnes, J. A., Loock, H. P., Saad, M., et al. (2014). Quantification of different water species in acetone using a NIR-triple-wavelength fiber laser. *Opt. Express* 22, 19337–19347. doi: 10.1364/OE.22.019337

Atanassova, S. (2015). Near infrared spectroscopy and aquaphotomics for monitoring changes during yellow cheese ripening. *Agric. Sci. Tech.* 7, 269–272.

Atanassova, S., Tsenkova, R., Vasu, R. M., Koleva, M., and Dimitrov, M. (2009). Identification of mastitis pathogens in raw milk by near infrared spectroscopy and SIMCA classification method. *Sci. Works Univ. Food Tech. Plovdiv* 56, 567–572.

Barnes, R., Dhanoa, M., and Lister, S. (1993). Correction to the description of standard normal variate (SNV) and de-trend (DT) transformations in practical spectroscopy with applications in food and beverage analysis-−2nd edition. *J. Near Infrared Spectrosc.* 1, 185–186. doi: 10.1255/jnirs.21

Barnes, R., Dhanoa, M. S., and Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 43, 772–777. doi: 10.1366/0003702894202201

Barzaghi, S., Cremonesi, K., and Cattaneo, T. M. P. (2017). Influence of the presence of bioactive compounds in smart-packaging materials on water absorption using NIR spectroscopy and aquaphotomics. *NIR News* 28, 21–24. doi: 10.1177/0960336017695731

Bázár, G., Kovacs, Z., Tanaka, M., Furukawa, A., Nagai, A., Osawa, M., et al. (2015). Water revealed as molecular mirror when measuring low concentrations of sugar with near infrared light. *Anal. Chim. Acta* 896, 52–62. doi: 10.1016/j.aca.2015.09.014

Bázár, G., Kovács, Z., Tanaka, M., and Tsenkova, R. (2014). "Aquaphotomics and its extended water mirror concept explain why NIRS can measure low concentration aqueous solutions," in *Aquaphotomics,"Understanding Water in Biological World". The 5th Kobe University Brussels European Centre Symposium "Innovation, Environment, and Globalisation"* (Brussels).

Bazar, G., Kovacs, Z., and Tsenkova, R. (2016). Evaluating spectral signals to identify spectral error. *PLoS ONE* 11:e0146249. doi: 10.1371/journal.pone.0146249

Bázár, G., Romvári, R., Szabó A., Somogyi, T., Éles, V., and Tsenkova R (2016). NIR detection of honey adulteration reveals differences in water spectral pattern. *Food Chem.* 194(Suppl. C), 873–880. doi: 10.1016/j.foodchem.2015.08.092

Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Sys.* 38, 149–171. doi: 10.1016/S0169-7439(97)00032-4

Buijs, K., and Choppin, G. R. (1963). Near-infrared studies of the structure of water. I. pure water. *J. Chem. Phys.* 39, 2035–2041. doi: 10.1063/1.1734579

Büning-Pfaue, H. (2003). Analysis of water in food by near infrared spectroscopy. *Food Chem.* 82, 107–115. doi: 10.1016/S0308-8146(02)00583-6

Cattaneo, T. M., Vanoli, M., Grassi, M., Rizzolo, A., and Barzaghi, S. (2016). The aquaphotomics approach as a tool for studying the influence of food coating materials on cheese and winter melon samples. *J. Near Infrared Spectrosc.* 24, 381–390. doi: 10.1255/jnirs.1238

Cattaneo, T. M. P., Cabassi, G., Profaizer, M., and Giangiacomo, R. (2009). Contribution of light scattering to near infrared absorption in milk. *J. Near Infrared Spectrosc.* 17, 337–343. doi: 10.1255/jnirs.867

Cattaneo, T. M. P., Stefania, V., Elena, N., and Vittorio, E. (2011). Influence of filtration processes on aqueous nanostructures by NIR spectroscopy. *J. Chem. Chem. Eng.* 5, 1046–1052.

Chandler, D. (2002). Hydrophobicity: two faces of water. *Nature* 417:491. doi: 10.1038/417491a

Chang, K., Shinzawa, H., and Chung, H. (2018). Concentration determination of inorganic acids that do not absorb near-infrared (NIR) radiation through recognizing perturbed NIR water bands by them and investigation of accuracy dependency on their acidities. *Microchem. J.* 139, 443–449. doi: 10.1016/j.microc.2018.03.019

Chatani, E., Tsuchisaka, Y., Masuda, Y., and Tsenkova, R. (2014). Water molecular system dynamics associated with amyloidogenic nucleation as revealed by real time near infrared spectroscopy and aquaphotomics. *PLoS ONE* 9:e101997. doi: 10.1371/journal.pone.0101997

Ciurczak, E. W., and Igne, B. (2014). *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*. Boca Raton, FL: CRC Press.

Cowe, I. A., and McNicol, J. W. (1985). The use of principal components in the analysis of near-infrared spectra. *Appl. Spectrosc.* 39, 257–266. doi: 10.1366/0003702854248944

Cui, X., Cai, W., and Shao, X. (2016). Glucose induced variation of water structure from temperature dependent near infrared spectra. *RSC Adv.* 6, 105729–105736. doi: 10.1039/C6RA18912A

Cui, X., Liu, X., Yu, X., Cai, W., and Shao, X. (2017a). Water can be a probe for sensing glucose in aqueous solutions by temperature dependent near infrared spectra. *Anal. Chim. Acta* 957, 47–54. doi: 10.1016/j.aca.2017.01.004

Cui, X., Zhang, J., Cai, W., and Shao, X. (2017b). Chemometric algorithms for analyzing high dimensional temperature dependent near infrared spectra. *Chemometr. Intell. Lab. Sys.* 170, 109–117. doi: 10.1016/j.chemolab.2017.08.010

Cupane, A., Levantino, M., and Santangelo, M. G. (2002). Near-infrared spectra of water confined in silica hydrogels in the temperature interval 365–5 K. *J. Phys. Chem B* 106, 11323–11328. doi: 10.1021/jp026117m

Czarnecki, M. A., Morisawa, Y., Futami, Y., and Ozaki, Y. (2015). Advances in molecular structure and interaction studies using near-infrared spectroscopy. *Chem. Rev.* 115, 9707–9744. doi: 10.1021/cr500013u

Czarnik-Matusewicz, B., Murayama, K., Tsenkova, R., and Ozaki, Y. (1999). Analysis of near-infrared spectra of complicated biological fluids by two-dimensional correlation spectroscopy: protein and fat concentration-dependent spectral changes of milk. *Appl. Spectrosc.* 53, 1582–1594. doi: 10.1366/0003702991946046

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap methods and their applications*, *Cambridge Series in Statistical and Probabilistic Mathematics*. Vol. 32. Cambridge, UK: Cambridge University Press.

Dhanoa, M., Lister, S., Sanderson, R., and Barnes, R. (1994). The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra. *J. Near Infrared Spectrosc.* 2, 43–47. doi: 10.1255/jnirs.30

Doster, W., Bachleitner, A., Dunau, R., Hiebl, M., and Lüscher, E. (1986). Thermal properties of water in myoglobin crystals and solutions at subzero temperatures. *Biophys. J.* 50, 213–219. doi: 10.1016/S0006-3495(86)83455-5

Fornés, V., and Chaussidon, J. (1978). An interpretation of the evolution with temperature of the $\nu 2+\nu 3$ combination band in water. *J. Chem. Phys.* 68, 4667–4671. doi: 10.1063/1.435576

Giangiacomo, R., Pani, P., and Barzaghi, S. (2009). Sugars as a perturbation of the water matrix. *J. Near Infrared Spectrosc.* 17, 329–335. doi: 10.1255/jnirs.861

Goto, N., Bazar, G., Kovacs, Z., Kunisada, M., Morita, H., Kizaki, S., et al. (2015). Detection of UV-induced cyclobutane pyrimidine dimers by near-infrared spectroscopy and aquaphotomics. *Scient. Rep.* 5:11808. doi: 10.1038/srep11808

Gowen, A. (2012). Water and food quality. *Contemp. Mater.* 1, 31–37. doi: 10.7251/COM1201031G

Gowen, A., Tsenkova, R., Esquerre, C., Downey, G., and O'Donnell, C. (2009a). Use of near infrared hyperspectral imaging to identify water matrix coordinates in mushrooms (Agaricus Bisporus) subjected to mechanical vibration. *J. Near Infrared Spectrosc.* 17, 363–371.

Gowen, A., Tsuchisaka, Y., O'Donnell, C., and Tsenkova, R. (2011). Investigation of the potential of near infrared spectroscopy for the detection and

quantification of pesticides in aqueous solution. *Am. J. Anal. Chem.* 2, 53–62. doi: 10.4236/ajac.2011.228124

Gowen, A. A., and Amigo, J. M. (2012). "Applications of spectroscopy and chemical imaging in pharmaceutics," in *Handbook of Biophotonics. Vol.3*: *Photonics in Pharmaceutics, Bionalysis and Environmental Research* eds J., Popp, V. V., Tuchin, A., Chiou, and, S., Heinemann (Weinheim: Wiley-VCH Verlag GmbHandCo), 71–88.

Gowen, A. A., Amigo, J. M., and Tsenkova, R. (2013). Characterisation of hydrogen bond perturbations in aqueous systems using aquaphotomics and multivariate curve resolution-alternating least squares. *Anal. Chim. Acta* 759, 8–20. doi: 10.1016/j.aca.2012.10.007

Gowen, A. A., Marini, F., Tsuchisaka, Y., De Luca, S., Bevilacqua, M., O'Donnell, C., et al. (2015). On the feasibility of near infrared spectroscopy to detect contaminants in water using single salt solutions as model systems. *Talanta* 131, 609–618. doi: 10.1016/j.talanta.2014.08.049

Gowen, A. A., Tsenkova, R., Esquerre, C., Downey, G., and O'Donnell, C. P. (2009b). Use of near infrared hyperspectral imaging to identify water matrix co-ordinates in mushrooms (Agaricus bisporus) subjected to mechanical vibration. *J. Near Infrared Spectrosc.* 17, 363–371. doi: 10.1255/jnirs.860

Grant, A., Davies, A. M. C., and Bilverstone, T. (1989). Simultaneous determination of sodium hydroxide, sodium carbonate and sodium chloride concentrations in aqueous solutions by near-infrared spectrometry. *Analyst* 114, 819–822. doi: 10.1039/an9891400819

Heiman, A., and Licht, S. (1999). Fundamental baseline variations in aqueous near-infrared analysis. *Anal. Chim. Acta* 394, 135–147. doi: 10.1016/S0003-2670(99)00312-8

Heise, H. M., Winzen, R. (2002)."Chemometrics in Near-Infrared Spectroscopy," in *Near-infrared spectroscopy: Principles, instruments and applications,* eds H. W., Siesler, Y., Ozaki, S., Kawata, and, H. M., Heise (Weinheim: Wiley-VCH Verlag GmbH), 125–162.

Hirschfeld, T. (1985). Salinity determination using NIRA. *Appl. Spectrosc.* 39, 740–741. doi: 10.1366/0003702854250293

Iwamoto, M., Uozumi, J., and Nishinari, K. (1987). "Preliminary investigation of the state of water in foods by near infrared spectroscopy," in *Proceedings of the International NIR/NIT Conference*. eds J. Hollo, K.J. Kafka and J. Gonzcy (Budapest: Akademiai Kiado).

Jamrógiewicz, M. (2012). Application of the near-infrared spectroscopy in the pharmaceutical technology. *J. Pharm. Biomed. Anal.* 66, 1–10. doi: 10.1016/j.jpba.2012.03.009

Jinendra, B. (2011). *Near Infrared Spectroscopy and Aquaphotomics: Novel Tool for Biotic and Abiotic Stress Diagnosis of Soybean.* Kobe: Kobe University.

Jinendra, B., Tamaki, K., Kuroki, S., Vassileva, M., Yoshida, S., and Tsenkova, R. (2010). Near infrared spectroscopy and aquaphotomics: novel approach for rapid in vivo diagnosis of virus infected soybean. *Biochem. Biophys. Res. Commun.* 397, 685–690. doi: 10.1016/j.bbrc.2010.06.007

Kang, J., Cai, W., and Shao, X. (2011). Quantitative determination by temperature dependent near-infrared spectra: a further study. *Talanta* 85, 420–424. doi: 10.1016/j.talanta.2011.03.089

Kinoshita, K., Kuze, N., Kobayashi, T., Miyakawa, E., Narita, H., Inoue-Murayama, M., et al. (2016). Detection of urinary estrogen conjugates and creatinine using near infrared spectroscopy in Bornean orangutans (Pongo pygmaeus). *Primates* 57, 51–59. doi: 10.1007/s10329-015-0501-3

Kinoshita, K., Miyazaki, M., Morita, H., Vassileva, M., Tang, C., Li, D., et al. (2012). Spectral pattern of urinary water as a biomarker of estrus in the giant panda. Scientific reports 2. doi: 10.1038/srep00856

Kinoshita, K., Morita, H., Miyazaki, M., Hama, N., Kanemitsu, H., Kawakami, H., et al. (2010). Near infrared spectroscopy of urine proves useful for estimating ovulation in giant panda (Ailuropoda melanoleuca). *Analytical Methods* 2, 1671–1675. doi: 10.1039/c0ay00333f

Kojić, D., Tsenkova, R., and Yasui, M. (2017). Improving accuracy and reproducibility of vibrational spectra for diluted solutions. *Anal. Chim. Acta* 955, 86–97. doi: 10.1016/j.aca.2016.12.019

Kojić, D., Tsenkova, R., Tomobe, K., Yasuoka, K., and Yasui, M. (2014). Water confined in the local field of ions. *Chemphyschem* 15, 4077–4086. doi: 10.1002/cphc.201402381

Kovacs, Z., Bázár, G., Oshima, M., Shigeoka, S., Tanaka, M., Furukawa, A., et al. (2016). Water spectral pattern as holistic marker for water quality monitoring. *Talanta* 147, 598–608. doi: 10.1016/j.talanta.2015.10.024

Liu, Y., Ozaki, Y., and Noda, I. (1996). Two-dimensional Fourier-transform near-infrared correlation spectroscopy study of dissociation of hydrogen-bonded N-methylacetamide in the pure liquid state. *J. Phys. Chem.* 100, 7326–7332. doi: 10.1021/jp9534186

Luck, W. A. (ed.). (1974). *Structure of Water and Aqueous Solutions.* Weinheim: Verlag Chemie.

Luck, W. A. P. (1998). The importance of cooperativity for the properties of liquid water. *J. Mol. Struct.* 448, 131–142. doi: 10.1016/S0022-2860(98)00343-3

Maeda, H., Ozaki, Y., Tanaka, M., Hayashi, N., and Kojima, T. (1995). Near Infrared spectroscopy and chemometrics studies of temperature-dependent spectral variations of water: relationship between spectral changes and hydrogen bonds. *J. Near Infrared Spectrosc.* 3, 191–201. doi: 10.1255/jnirs.69

Manley, M. (2014). Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chem. Soc. Rev.* 43, 8200–8214. doi: 10.1039/C4CS00062E

Martens, H., and Martens, M. (2001). "Multivariate analysis of quality. An introduction." (Chichester: Wiley).

Matija, L., Muncan, J., Mileusnic, I., and Koruga, D. "Fibonacci nanostructures for novel nanotherapeutical approach," in Nano-and Microscale Drug Delivery, S.ystems., Elsevier,(2017). 49-74. doi: 10.1016/B978-0-323-52727-9.00004-2

Matija, L., Tsenkova, R., Miyazaki, M., Banba, K., and Muncan, J. (2012). Aquagrams: Water spectral pattern as characterization of hydrogenated nanomaterial. *FME Transac.* 40, 51–56.

Matija, L., Tsenkova, R., Munćan, J., Miyazaki, M., Banba, K., Tomić M., et al. (2013). "Fullerene based nanomaterials for biomedical applications: engineering, functionalization and characterization," in: *Advanced Materials Research*: Trans Tech Publ, 224–238. doi: 10.4028/www.scientific.net/AMR.633.224

Meilina, H., Kuroki, S., Jinendra, B. M., Ikuta, K., and Tsenkova, R. (2009). Double threshold method for mastitis diagnosis based on NIR spectra of raw milk and chemometrics. *Biosyst. Eng.* 104, 243–249. doi: 10.1016/j.biosystemseng.2009.04.006

Meilina, H., Putra, A., and Tsenkova, R. (2011). "Frequency of use minute concentrations of cadmium in aqueous solution by near infrared spectroscopy and aquaphotomics," in *Proceedings of the Annual International Conference*, Syiah Kuala University-Life Sciences & Engineering Chapter.

Monakhova, Y. B., Pozharov, M. V., Zakharova, T. V., Khvorostova, E. K., Markin, A. V., Lachenmeier, D. W., et al. (2014). Association/Hydrogen bonding of acetone in polar and non-polar solvents: NMR and NIR spectroscopic investigations with chemometrics. *J. Solution Chem.* 43, 1963–1980. doi: 10.1007/s10953-014-0249-1

Munćan, J., Matija, L., Simić-Krstić J., Nijemčević S., and Koruga, D. (2014). Discrimination of mineral waters using near infrared spectroscopy and aquaphotomics. *Hemijska industrija* 68, 257–264. doi: 10.2298/HEMIND130412049M

Munćan, J., Mileusnić I., Matović V., Šakota Rosić J., and Matija, L. (2016a). *"The prospects of aquaphotomics in biomedical science and engineering,"* in Aquaphotomics: Understanding Water in Biology – 2nd International Symposium. (Kobe University, Kobe, Japan).

Munćan, J., Mileusnić I., Šakota Rosić J., Vasić-Milovanović A., and Matija, L. (2016b). Water properties of soft contact lenses: a comparative near-infrared study of two hydrogel materials. *Int. J. Polym. Sci.*1–8. doi: 10.1155/2016/3737916

Murayama, K., Czarnik-Matusewicz, B., Wu, Y., Tsenkova, R., and Ozaki, Y. (2000). Comparison between conventional spectral analysis methods, chemometrics, and two-dimensional correlation spectroscopy in the analysis of near-infrared spectra of protein. *Appl. Spectrosc.* 54, 978–985. doi: 10.1366/0003702001950715

Murayama, K., Yamada, K., Tsenkova, R., Wang, Y., and Ozaki, Y. (1998). Near-infrared spectra of serum albumin and γ-globulin and determination of their concentrations in phosphate buffer solutions by partial least squares regression. *Vib. Spectrosc.* 18, 33–40. doi: 10.1016/S0924-2031(98)00034-4

Næs, T., Isaksson, T., Fearn, T., Davies, T. A., (2002). *User Friendly Guide to Multivariate Calibration and Classification*. (Chichester: NIR publications).

Nakakimura, Y., Vassileva, M., Stoyanchev, T., Nakai, K., Osawa, R., Kawano, J., et al. (2012). Extracellular metabolites play a dominant role in near-infrared spectroscopic quantification of bacteria at food-safety level concentrations. *Anal. Methods* 4, 1389–1394. doi: 10.1039/c2ay05771a

Noda, I., Liu, Y., Ozaki, Y., and Czarnecki, M. A. (1995). Two-dimensional Fourier transform near-infrared correlation spectroscopy studies of temperature-dependent spectral variations of oleyl alcohol. *J. Phys. Chem.* 99, 3068–3073. doi: 10.1021/j100010a016

Noda, I., and Ozaki, Y. (eds.). (2004). *Two-Dimensional Correlation Spectroscopy – Applications in Vibrational and Optical Spectroscopy*. Chichester: John Wiley and Sons Inc.

Norris, K. H., and Williams, P. C. (1984). Optimization of mathematical treatments of raw near-infrared signal in the. measurement of protein in hard red spring wheat. i. influence of particle size**.** *Cereal Chem.* 61, 158–165.

Omar, A. F., Atan, H., and Matjafri, M. Z. (2012). NIR spectroscopic properties of aqueous acids solutions. *Molecules* 17, 7440–7450. doi: 10.3390/molecules17067440

Osborne, B. G., Fearn, T., Hindle, P. H., and Practical, N. I. R., (1993). *Spectroscopy with Applications in Food and Beverage Analysis*. Harlow: Longman scientific and technical.

Ozaki, Y. (2002) "Applications in Chemistry," in *Near-infrared Spectroscopy: Principles, Instruments, Applications,* eds H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise (Weinheim: Verlag GmbH), 179–212.

Ozaki, Y., Katsumoto, Y., Jiang, J.-H., Liang, Y. (2003) "Spectral analysis in the, NIR region" in *Useful and Advanced Information in the Field of near Infrared Spectroscopy,* ed S. Tsuchikawa (Trivandrum: Research Signpost), 307.

Pasquini, C. (2018). Near infrared spectroscopy: a mature analytical technique with new perspectives – A review. *Anal. Chim. Acta*. 1026, 8–36. doi: 10.1016/j.aca.2018.04.004

Patil, R. (2015). Noise reduction using wavelet transform and singular vector decomposition. *Procedia Comput. Sci.* 54, 849–853. doi: 10.1016/j.procs.2015.06.099

Peinado, A. C., van den Berg, F., Blanco, M., and Bro, R. (2006). Temperature-induced variation for NIR tensor-based calibration. *Chemometr. Intell. Lab. Sys.* 83, 75–82. doi: 10.1016/j.chemolab.2006.01.006

Pollner, B., and Kovacs, Z. (2016). Multivariate data analysis tools for, R including aquaphotomics methods, aquap2

Putra, A., Faridah, F., Inokuma, E., and Santo, R. (2010). Robust spectral model for low metal concentration measurement in aqueous solution reveals the importance of water absorbance bands. *J. Sains dan Teknologi Reaksi* 210:8.

Putra, A., Vassileva, M., Santo, R., and Tsenkova, R. (2017). "An efficient near infrared spectroscopy based on aquaphotomics technique for rapid determining the level of Cadmium in aqueous solution," in *IOP Conference Series: Materials Science and Engineering* (Kuala Lumpur).

R Core Team(2017). *A language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reeves, J. B. (1995). Efforts to quantify changes in near-infrared spectra caused by the influence of water, pH, ionic strength, and differences in physical state. *Appl. Spectrosc.* 49, 181–187. doi: 10.1366/0003702953963788

Remagni, M. C., Morita, H., Koshiba, H., Cattaneo, T. M. P., and Tsenkova, R. (2013). "Near infrared spectroscopy and aquaphotomics as tools for bacteria classification. NIR*2013,* in *Proceedings: Picking Up Good Vibrations* (La Grande-Motte), 602.

Rinnan, Å., van den Berg, F., and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends Anal. Chem.* 28, 1201–1222. doi: 10.1016/j.trac.2009.07.007

Robertson, W. H., Diken, E. G., Price, E. A., Shin, J.-W., and Johnson, M. A. (2003). Spectroscopic determination of the OH– solvation shell in the OH–·(H2O) n clusters. *Science* 299, 1367–1372. doi: 10.1126/science.1080695

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., and Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed. Anal.* 44, 683–700. doi: 10.1016/j.jpba.2007.03.023

Šakota Rosić, J., Munćan, J., Mileusnić I., Kosić B., and Matija, L. (2016). Detection of protein deposits using NIR spectroscopy. *Soft Mater.* 14, 264–271. doi: 10.1080/1539445X.2016.1198377

Sakudo, A., Tsenkova, R., Onozuka, T., Morita, K., Li, S., Warachit, J., et al. (2005). A novel diagnostic method for human immunodeficiency virus Type-1 in plasma by near-infrared spectroscopy. *Microbiol. Immunol.* 49, 695–701. doi: 10.1111/j.1348-0421.2005.tb03648.x

Sakudo, A., Tsenkova, R., Tei, K., Morita, H., Ikuta, K., and Onodera, T. (2006a). *Ex vivo* tissue discrimination by visible and near-infrared spectra with chemometrics. *J. Vet. Med. Sci.* 68, 1375–1378. doi: 10.1292/jvms.68.1375

Sakudo, A., Tsenkova, R., Tei, K., Onozuka, T., Ikuta, K., Yoshimura, E., et al. (2006b). Comparison of the vibration mode of metals in HNO3 by a partial least-squares regression analysis of near-infrared spectra. *Biosci. Biotechnol. Biochem.* 70, 1578–1583. doi: 10.1271/bbb.50619

Sakudo, A., Yoshimura, E., Tsenkova, R., Ikuta, K., and Onodera, T. (2007). Native state of metals in non-digested tissues by partial least squares regression analysis of visible and near-infrared spectra. *J. Toxicol. Sci.* 32, 135–141. doi: 10.2131/jts.32.135

Sartor, G., Hallbrucker, A., and Mayer, E. (1995). Characterizing the secondary hydration shell on hydrated myoglobin, hemoglobin, and lysozyme powders by its vitrification behavior on cooling and its calorimetric glass–>liquid transition and crystallization behavior on reheating. *Biophys. J.* 69, 2679–2694.

Šašić, S., Segtnan, V. H., and Ozaki Y (2002). Self-modeling curve resolution study of temperature-dependent near-infrared spectra of water and the investigation of water structure. *J. Phys. Chem. A* 106, 760–766. doi: 10.1021/jp013436p

Savitzky, A., and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639. doi: 10.1021/ac60214a047

Segtnan, V. H., Sasić S., Isaksson T, and Ozaki, Y. (2001). Studies on the structure of water using two-dimensional near-infrared correlation spectroscopy and principal component analysis. *Anal. Chem.* 73, 3153–3161. doi: 10.1021/ac010102n

Shan, R., Zhao, Y., Fan, M., Liu, X., Cai, W., and Shao, X. (2015). Multilevel analysis of temperature dependent near-infrared spectra. *Talanta* 131, 170–174. doi: 10.1016/j.talanta.2014.07.081

Shao, X., Cui, X., Yu, X., and Cai, W. (2018). Mutual factor analysis for quantitative analysis by temperature dependent near infrared spectra. *Talanta* 183, 142–148. doi: 10.1016/j.talanta.2018.02.043

Shao, X., Kang, J., and Cai, W. (2010). Quantitative determination by temperature dependent near-infrared spectra. *Talanta* 82, 1017–1021. doi: 10.1016/j.talanta.2010.06.009

Shao, X. G., Leung, A. K., and Chau, F. T. (2003). Wavelet: a new trend in chemistry. *Acc. Chem. Res.* 36, 276–283. doi: 10.1021/ar990163w

Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M. (2008). *Near-Infrared Spectroscopy: Principles, Instruments, Applications.* Weinheim: John, Wiley and Sons.

Slavchev, A., Kovacs, Z., Koshiba, H., Bazar, G., Pollner, B., Krastanov, A., et al. (2017). Monitoring of water spectral patterns of lactobacilli development as a tool for rapid selection of probiotic candidates. *J. Near Infrared Spectrosc.* 25:0967033517741133. doi: 10.1177/0967033517741133

Slavchev, A., Kovacs, Z., Koshiba, H., Nagai, A., Bázár, G., Krastanov, A., et al. (2015). Monitoring of water spectral pattern reveals differences in probiotics growth when used for rapid bacteria selection. *PLoS ONE* 10:e0130698. doi: 10.1371/journal.pone.0130698

Smith, J. D., Cappa, C. D., Wilson, K. R., Cohen, R. C., Geissler, P. L., and Saykally, R. J. (2005). Unified description of temperature-dependent hydrogen-bond rearrangements in liquid water. *Proc. Natl. Acad. Sci. U.S.A.* 102, 14171–14174. doi: 10.1073/pnas.0506899102

Steen, G. W., Fuchs, E. C., Wexler, A. D., and Offerhaus, H. L. (2015). Identification and quantification of 16 inorganic ions in water by Gaussian curve fitting of near-infrared difference absorbance spectra. *Appl. Opt.* 54, 5937–5942. doi: 10.1364/AO.54.005937

Takemura, G., Bázár, G., Ikuta, K., Yamaguchi, E., Ishikawa, S., Furukawa, A., et al. (2015). Aquagrams of raw milk for oestrus detection in dairy cows. *Reprod. Domest. Anim.* 50, 522–525. doi: 10.1111/rda.12504

Tillmann, P., and Paul, C. (1998). The repeatability file—a tool for reducing the sensitivity of near infrared spectroscopy calibrations to moisture variation. *J. Near Infrared Spectrosc.* 6, 61–68. doi: 10.1255/jnirs.122

Timmerman, M. E. (2006). Multilevel component analysis. *Br. J. Math. Stat. Psychol.* 59, 301–320. doi: 10.1348/000711005X67599

Tsenkova, R. (2004). *Near Infrared Spectroscopy of Raw Milk for Cow's Biomonitoring*. Ph.D. thesis, Hokkaido University (?????).

Tsenkova, R. (2005). "Visible-near infrared perturbation spectroscopy: Water in action seen as a source of information," in *12th International Conference on Near-infrared Spectroscopy* (Auckland), 607–612.

Tsenkova, R. (2006a). Aquaphotomics. Aquaphotomics and chambersburg. *NIR News* 17, 12–10. doi: 10.1255/nirn.916

Tsenkova, R. (2006b). Aquaphotomics: exploring water-light interactions for a better understanding of the biological world. Part 2: Japanese food, language and why NIR for diagnosis? *NIR News* 17, 8–14. doi: 10.1255/nirn.904

Tsenkova, R. (2006c). AquaPhotomics: water absorbance pattern as a biological marker. *NIR News* 17, 13–10. doi: 10.1255/nirn.1014

Tsenkova, R. (2007). AquaPhotomics: water absorbance pattern as a biological marker for disease diagnosis and disease understanding. *NIR News* 18, 14–16. doi: 10.1255/nirn.1014

Tsenkova, R. (2008a). Aquaphotomics: acquiring spectra of various biological fluids of the same organism reveals the importance of water matrix absorbance coordinates and the aquaphotome for understanding biological phenomena. *NIR News* 19, 13–15.

Tsenkova, R. (2008b). Aquaphotomics: the extended water mirror effect explains why small concentrations of protein in solution can be measured with near infrared light. *NIR News* 19, 13–14.

Tsenkova, R. (2008c). "Aquaphotomics: VIS-near infrared spectrum of water as biological marker," in *Conference on the Physics, Chemistry and Biology of Water* (Sofia).

Tsenkova, R. (2009). Aquaphotomics: dynamic spectroscopy of aqueous and biological systems describes peculiarities of water. *J. Near Infrared Spectrosc.* 17, 303–313. doi: 10.1255/jnirs.869

Tsenkova, R. (2010). Aquaphotomics: water in the biological and aqueous world scrutinised with invisible light. *Spectrosc. Eur.* 22, 6–10.

Tsenkova, R., and Atanassova, S. (2002). "Mastitis diagnostics by near infrared spectra of cow's milk, blood and urine using soft independent modelling of class analogy classification," in *Near Infrared Spectroscopy: Proceedings of the 10th International Conference,* eds A. M. C. Davies and R. K. Cho (Chichester: NIR Publications).

Tsenkova, R., Atanassova, S., Kawano, S., and Toyoda, K. (2001a). Somatic cell count determination in cow's milk by near-infrared spectroscopy: a new diagnostic tool. *J. Anim. Sci.* 79, 2550–2557. doi: 10.2527/2001.79102550x

Tsenkova, R., Atanassova, S., Ozaki, Y., Toyoda, K., and Itoh, K. (2001b). Near-infrared spectroscopy for biomonitoring: influence of somatic cell count on cow's milk composition analysis. *Intl. Dairy J.* 11, 779–783. doi: 10.1016/S0958-6946(01)00110-8

Tsenkova, R., Atanassova, S., and Toyoda, K. (2001c). Near infrared spectroscopy for diagnosis: influence of mammary gland inflammation on cow' s milk composition measurement. *Near Infrared Anal.* 2, 59–66. doi: 10.11357/jsam1937.61

Tsenkova, R., Fockenberg, C., Koseva, N., Sakudo, A., and Parker, M. (2007a). "Aquaphotomics: water absorbance patterns in NIR range used for detection of metal ions reveal the importance of sample preparation," in *13th International Conference on Near Infrared spectroscopy* (Umea), 03–02.

Tsenkova, R., Iso, E., Parker, M., Fockenberg, C., and Okubo, M. (2007b). "Aquaphotomics: a NIRS investigation into the perturbation of water spectrum in an aqueous suspension of mesoscopic scale polystyrene spheres," in *13th International Conference on Near Infrared Spectroscopy* (Umea), *A*–04.

Tsenkova, R., Kovacs, Z., Kubota, Y., (2015) "Aquaphotomics: near infrared spectroscopy and water states in biological systems," in *Membrane Hydration,* ed E. Anibal Disalvo (Berlin: Springer), 189–211.

Tsenkova, R., Morita, H., Shinzawa, H., Hogeveen, H., Hillerton, J. E., and Ikuta, K. (2005). "Near infrared spectroscopy for cow identification and *in-vivo* mastitis diagnosis," in *Mastitis in Dairy Production. Current Knowledge and Future Solutions, 4th IDF International Mastitis Conference* (Maastricht), 901.

Tsenkova, R. N. (1994). "Near-infrared spectroscopy of individual cow milk as a means for automated monitoring of udder health and milk quality," in *Proceedings of Third International Dairy Housing Conference* (Orlando, FL).

Tsenkova, R. N., Iordanova, I. K., Toyoda, K., and Brown, D. R. (2004). Prion protein fate governed by metal binding. *Biochem. Biophys. Res. Commun.* 325, 1005–1012. doi: 10.1016/j.bbrc.2004.10.135

Wang, Y., Murayama, K., Myojo, Y., Tsenkova, R., Hayashi, N., and Ozaki, Y. (1998). Two-dimensional fourier transform near-infrared spectroscopy study of heat denaturation of ovalbumin in aqueous solutions. *J. Phys. Chem. B* 102, 6655–6662. doi: 10.1021/jp9816115

Weber, J. M., Kelley, J. A., Nielsen, S. B., Ayotte, P., and Johnson, M. A. (2000). Isolating the spectroscopic signature of a hydration shell with the use of clusters: superoxide tetrahydrate. *Science* 287, 2461–2463. doi: 10.1126/science.287.5462.2461

Weber, J. M., Kelley, J. A., Robertson, W. H., and Johnson, M. A. (2001). Hydration of a structured excess charge distribution: infrared spectroscopy of the $O_2^- \cdot (H_2O)_n$, $(1 \leq n \leq 5)$ clusters. *J. Chem. Phys.* 114, 2698–2706. doi: 10.1063/1.1338529

Wenz, J. J. (2018). Examining water in model membranes by near infrared spectroscopy and multivariate analysis. *Biochim. Biophys. Acta Biomembr.* 1860, 673–682. doi: 10.1016/j.bbamem.2017.12.007

Williams, P., Norris, K. (1987). *Near-Infrared Technology in the Agricultural and Food Industries*. St. Paul, MI: American Association of Cereal Chemists Inc.

Wold, S., Geladi, P., Esbensen, K., and Öhman J. (1987). Multi-way principal components-and PLS-analysis. *J. Chemom.* 1, 41–56. doi: 10.1002/cem.1180010107

Wold, S., Sjöström, M. (1977). "SIMCA: a method for analyzing chemical data in terms of similarity and analogy," in *Chemometrics: Theory and Application,* ed B. R. Kowalski (Washington DC: American Chemical Society), 243–282.

Workman, J. Jr. (2000). *The Handbook of Organic Compounds: NIR, IR, Raman, and, UV-, VIS Spectra Featuring Polymers and Surfactants*. London: Elsevier.

Workman, J. (2016). *The Concise Handbook of Analytical Spectroscopy, Vol. 3*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Wu, H. L., Shibukawa, M., and Oguma, K. (1998). An alternating trilinear decomposition algorithm with application to calibration of HPLC–DAD for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. *J. Chemometr. Soc.* 12, 1–26.

Xantheas, S. S. (1995). *Ab initio* studies of cyclic water clusters $(H_2O)_n$, $n$=1–6. III. Comparison of density functional with MP2 results. *J. Chem. Phys.* 102, 4505–4517. doi: 10.1063/1.469499

Yuan, B., Murayama, K., Wu, Y., Tsenkova, R., Dou, X., Era, S., et al. (2003). Temperature-dependent near-infrared spectra of bovine serum albumin in aqueous solutions: spectral analysis by principal component analysis and evolving factor analysis. *Appl. Spectrosc.* 57, 1223–1229. doi: 10.1366/000370203769699072

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 10 September 2018
doi: 10.3389/fchem.2018.00400

# Raman Spectroscopy for Pharmaceutical Quantitative Analysis by Low-Rank Estimation

*Xiangyun Ma*[1], *Xueqing Sun*[1], *Huijie Wang*[1], *Yang Wang*[1], *Da Chen*[2] *and Qifeng Li*[1]*

[1] School of Precision Instrument and Opto-electronics Engineering, Tianjin University, Tianjin, China, [2] State Key Laboratory of Precision Measurement Technology and Instruments, Tianjin University, Tianjin, China

Raman spectroscopy has been widely used for quantitative analysis in biomedical and pharmaceutical applications. However, the signal-to-noise ratio (SNR) of Raman spectra is always poor due to weak Raman scattering. The noise in Raman spectral dataset will limit the accuracy of quantitative analysis. Because of high correlations in the spectral signatures, Raman spectra have the low-rank property, which can be used as a constraint to improve Raman spectral SNR. In this paper, a simple and feasible Raman spectroscopic analysis method by Low-Rank Estimation (LRE) is proposed. The Frank-Wolfe (FW) algorithm is applied in the LRE method to seek the optimal solution. The proposed method is used for the quantitative analysis of pharmaceutical mixtures. The accuracy and robustness of Partial Least Squares (PLS) and Support Vector Machine (SVM) chemometric models can be improved by the LRE method.

**Keywords: Raman spectroscopy, quantitative analysis, pharmaceuticals, low-rank estimation, chemometric model**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Andreas Borgschulte,*
*Swiss Federal Laboratories for Materials Science and Technology, Switzerland*
*Pellegrino Musto,*
*Consiglio Nazionale Delle Ricerche (CNR), Italy*

***Correspondence:**
*Qifeng Li*
*Lqfli@tju.edu.cn*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *28 February 2018*
***Accepted:*** *20 August 2018*
***Published:*** *10 September 2018*

***Citation:***
*Ma X, Sun X, Wang H, Wang Y, Chen D and Li Q (2018) Raman Spectroscopy for Pharmaceutical Quantitative Analysis by Low-Rank Estimation. Front. Chem. 6:400. doi: 10.3389/fchem.2018.00400*

## INTRUDUCTION

Raman spectroscopy is one of the vibrational spectroscopic techniques that has been commonly applied in quantitative analysis (Strachan et al., 2004; Numata and Tanaka, 2011; Ai et al., 2018). Being non-invasive and marker-free, it has been proved to be an effective tool in the field of physics, chemistry, and biology (Graf et al., 2007; Neugebauer et al., 2010; Ryu et al., 2012; Tan et al., 2017). Coupled with chemometrics methods, it has the advantages of high sensitivity and resolution in biomedical and pharmaceutical quantitative analysis.

The quantitative analysis based on Raman spectra at low signal-to-noise ratio (SNR) levels is still problematic (Li, 2008; Chen et al., 2014). Generally, a Raman spectrum can be divided into two parts: the signal containing desired information and the noise containing unwanted information. Basically, the latter may include photon-shot noise, sample-generated noise, instrument-generated noise, computationally generated noise, and externally generated noise (Pelletier, 2003). Due to the inherently weak property of Raman scattering, the noise will lead to a deterioration in SNR of Raman spectra, affecting the accuracy of quantitative analysis. For instance, data of online monitoring in limited integration time always tend to be inaccurate (Han et al., 2017; Virtanen et al., 2017).

Some approaches of preprocessing Raman spectra to minimize this problem have been proposed (Clupek et al., 2007; Ma et al., 2017), such as first and second derivatives (Johansson et al., 2010), polynomials fitting (Vickers et al., 2001), Fourier transform (Pelletier, 2003), and wavelet transform (Chen et al., 2011; Li et al., 2013). Among these approaches, wavelet transform can extract peak information and remove background noise, which has been the most widely used preprocessing method (Du et al., 2006). However, the processing of Raman spectra can be further optimized to improve the accuracy of pharmaceutical quantitative analysis.

In this paper, we introduce a simple and feasible Raman spectroscopic analysis method based on Low-Rank Estimation (LRE). Our experiments are implemented based on the Partial Least Squares (PLS) and Support Vector Machine (SVM) chemometric models. The aim of this experimental design is to enhance the quality of pharmaceutical quantitative analysis by significantly improving the accuracy and robustness of the chemometric models used.

## MATERIALS AND METHODS

Pharmaceutical substances (norfloxacin, penicillin potassium, and sulfamerazine) were purchased from Dalian Meilun Biotechnology Co., Ltd (China) and used without further purification. These substances were well blended in different proportions, pulverized, and compressed into three-component tablets. Other physical properties of these tablets (such as density, height, and diameter) were kept completely consistent. Mixed solutions were also prepared with methanol and ethanol in 100 different proportions. Raman spectral data were recorded by using a Renishaw inVia Raman spectrometer (Gloucestershire, U.K.). This system consisted of a 785-nm diode laser (∼40 mW) and a 1,200 l/mm grating. In this work, the integration times of Raman spectra were 0.1–0.5 s.

PLS and SVM regression methods were used to model and predict pharmaceutical concentration of the samples based on their Raman spectra. Eighty-five samples were selected as the training set and the remaining 15 samples as the testing set, based on Kennard-Stone (KS) algorithm. The parameters of PLS and SVM models were tuned based on grid search algorithm. The optimal parameters were obtained by k-folder cross-validation.

The accuracy and robustness of above-mentioned chemometric models were further improved by conventional Wavelet Transform (WT) method and Low-Rank Estimation (LRE) method, respectively. In the WT method, the signals were split into different frequency components to remove simultaneously low-frequency background and high-frequency noise components. The Symlet wavelet filter (sym11, scale = 7) was optimally selected to provide the sharpest peaks associated with the analytes of interest. The LRE method was originally developed by our group in three-dimension to speed up Raman spectral imaging (Li et al., 2018). In this study, we used the LRE method in two-dimension to process the observed Raman spectral data matrix. In this method, the alternating least squares (ALS) algorithm is used to estimate the largest singular value of the matrix (Kroonenberg and Leeuw, 1980; Halko et al., 2011). The matrix estimation has two sets of parameters. Each set is estimated in turn by solving a least-squares problem and holding the other set fixed. After both sets have been estimated once, the procedure is repeated until convergence.

The Frank-Wolfe (FW) algorithm is applied in the LRE method to seek the optimal solution. Recently, the FW algorithm has been popularly used in machine learning due to its characteristics of simple implementation and modest memory requirement (Jaggi, 2013; Guo et al., 2017). The steps of the LRE method are detailed in **Table 1**.

**TABLE 1 |** The detail steps of the LRE method.

**Algorithm**: The algorithm for the LRE method

**Input**: the raw Raman spectral data matrix $A$;
the maximum number of iteration $N$, ranging from 5 to 20;
the low-rank constraint factor $m$, ranging from 0.01 to 0.001;

1: **Initialize** $X^0 = 0$. $X^0$ is an initial solution of the algorithm.

2: **for** $i = 0,1,\ldots,N$ **do**, $a^i$ represents the $i$-th iteration of any variable $a$.

3: Compute the search direction $s$, $s^{i+1} = \mathrm{ALS}(A - X^i)$

4: Compute the step length $r$, $r^{i+1} = \arg\min_{r\in[0,1]}(A - (X^i + r(s^{i+1} - X^i)))$

5: $X^{i+1} = (1 - r^{i+1})X^i + r^{i+1}s^{i+1}$

6: **stopping criterion:** $\frac{\mathrm{ALS}(X^{i+1})}{s^{i+1}} > m$

7: **end for**

8: The last iteration of $X$ is the final solution of the LRE method.

**Output** $X$

Through being processed by the LRE method, the low-rank training and testing sets can be obtained from the raw training and testing data matrices, respectively. In general, an abundant data matrix can enhance the effect of the LRE method. When a number of testing spectral data is small, the training spectral data can be added to the raw testing data matrix as a supplement. The added spectral data are only used to strengthen the impact of the LRE method. The conventional regression models are applied to the low-rank training and testing sets to perform quantitative Raman analysis.

## RESULTS AND DISCUSSION

Noise-free Raman spectral dataset is a low-rank matrix. In **Figure 1**, the red line shows the ranks of Raman spectral data matrix in an integration time of 1 s, suggesting that the Raman spectra have low-rank property when the noise is low. The low-rank property comes from high correlations among spectral signatures. Each spectral signature can be represented by a linear combination of a small number of pure spectral endmembers, which is known as linear spectral mixing model (Iordache et al., 2011; Golbabaee and Vandergheynst, 2012). The blue and green plots show singular values of the matrix in a shorter integration time, which implies that the ranks of Raman spectra increase with decreasing integration time owing to a greater proportion of the noise. The low-rank property can be used as a constraint to improve the accuracy of pharmaceutical quantitative analysis (Yi et al., 2017).

Raw Raman spectra recorded for three pure pharmaceutical substances are shown in **Figure 2A**. Thirty Raman spectra obtained from three-component tablets with different proportions are shown in **Figure 2B**. It is clear that each pharmaceutical component has its own special characteristic peaks. However, their respective Raman bands are overlapped. Particularly, Raman signals of lower-concentration component are almost swamped and covered by those of higher-concentration one, which represents a common problem in practice for biomedical and pharmaceutical quantitative analysis. For clarity, the Raman spectra in **Figure 2B** were collected

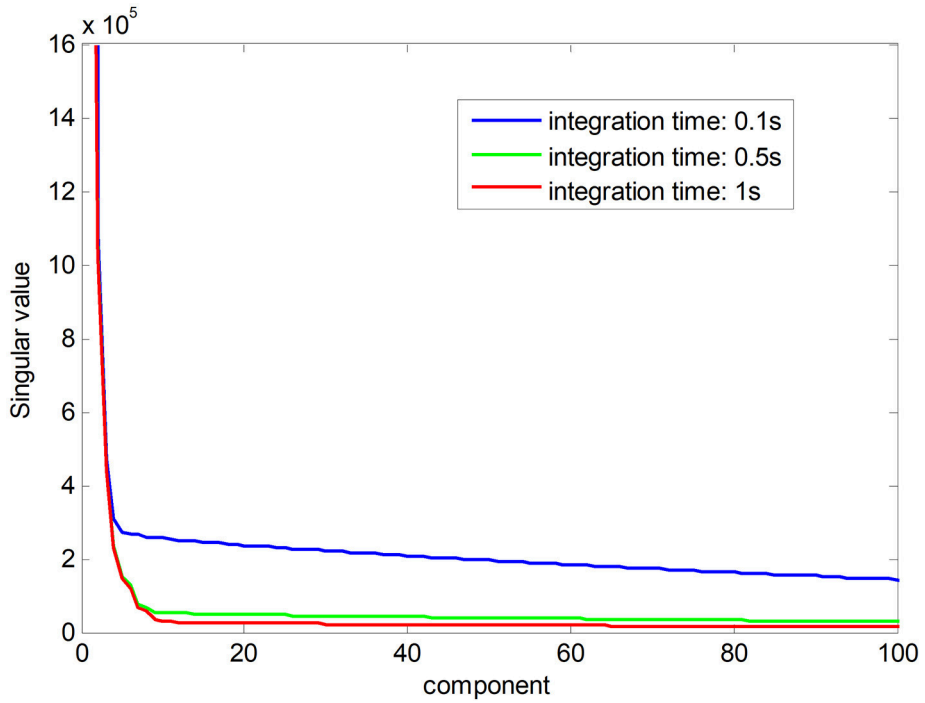



**FIGURE 1 |** The ranks of the Raman spectra in different integration time.

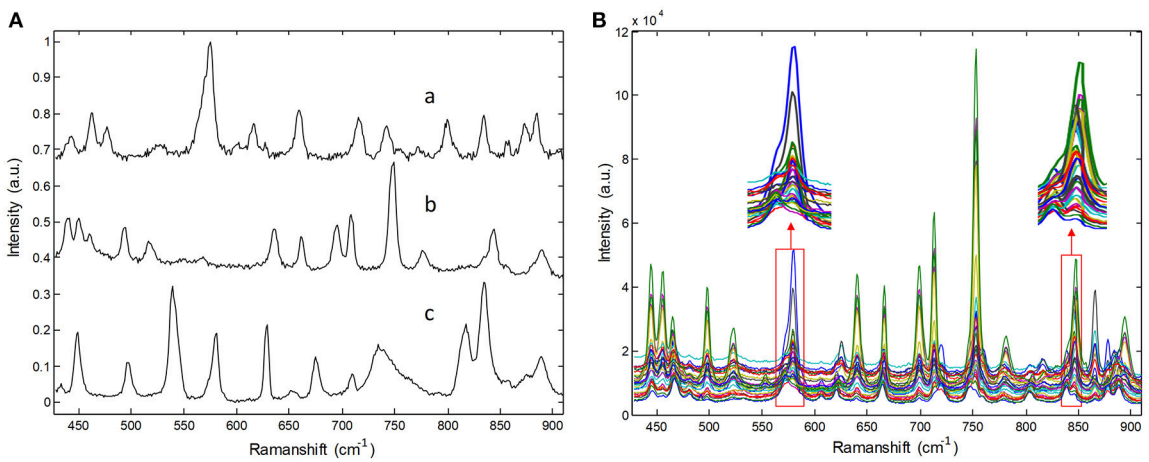



**FIGURE 2 |** Raman spectra of **(A)** pure pharmaceuticals substances: (a) penicillin potassium, (b) norfloxacin, and (c) sulfamerazine, **(B)** their three component tablets containing different proportions in the integration time of 5s.

in an integration time of 5 s, which have a high SNR. In our experiments, the integration times of Raman spectra are in the range of 0.1–0.5 s, which is over 10 times shorter than that shown in **Figure 2**. Under this condition, the spectral signals are weaker and have poor SNR.

The comparisons of predicted and actual values for norfloxacin are illustrated in **Figure 3**, which indicates the advantage of the LRE method for pharmaceutical quantitative analysis. The coefficient of determination ($R^2$) and root mean square error (RMSE) of the chemometric models used for quantitative analysis of three pharmaceutical components are listed in **Table 2**. The unsatisfactory results of the raw spectral data show that the pre-treatment of Raman spectra is necessary. In this study, the LRE method and conventional wavelet transform (WT) method are applied to improve the accuracy of quantitative analysis. As shown in **Figure 3**, both the conventional WT and LRE methods can improve the predicted results. However, it is clear that the LRE method has a better

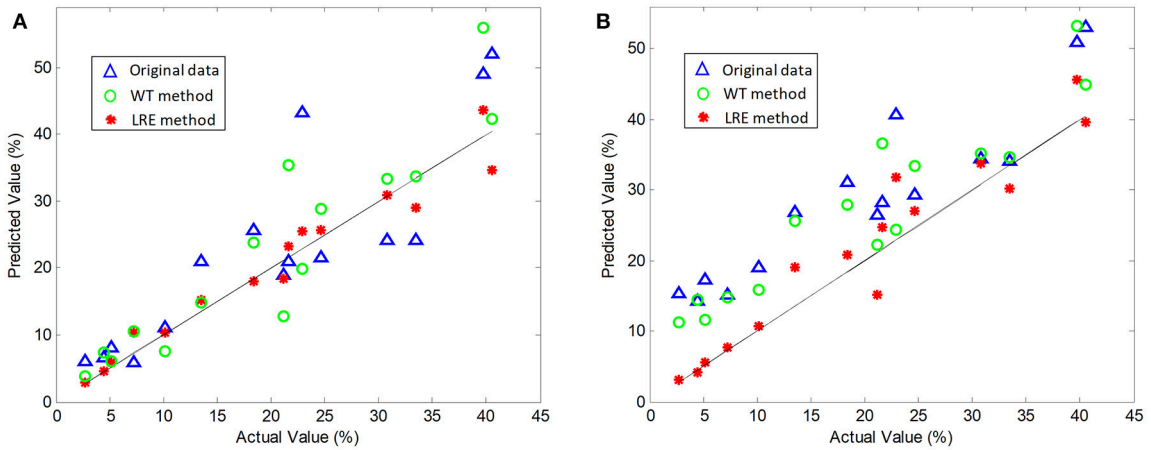



**FIGURE 3 |** Actual vs. predicted values of norfloxacin based on the PLS **(A)** and SVM **(B)** model, where the black solid line are diagonals. Raw Raman spectra are collected in an integration time of 0.2 s.

**TABLE 2 |** $R^2$ and RMSE values of the chemometric models for three pharmaceutical components.

| | | Norfloxacin | | Penicillin potassium | | Sulfamerazine | |
|---|---|---|---|---|---|---|---|
| **Model** | **Methods** | $R^2$ | **RMSE** | $R^2$ | **RMSE** | $R^2$ | **RMSE** |
| **PLS** | Raw | 0.7504 | 0.0780 | 0.8692 | 0.1218 | 0.7323 | 0.0608 |
| | WT | 0.8598 | 0.0642 | 0.9548 | 0.0974 | 0.8862 | 0.0376 |
| | LRE | 0.9553 | 0.0259 | 0.9848 | 0.0522 | 0.9609 | 0.0225 |
| **SVM** | Raw | 0.8297 | 0.1097 | 0.8460 | 0.1264 | 0.8135 | 0.0679 |
| | WT | 0.8808 | 0.0841 | 0.9125 | 0.0821 | 0.8907 | 0.0444 |
| | LRE | 0.9558 | 0.0468 | 0.9749 | 0.0755 | 0.9701 | 0.0397 |

**TABLE 3 |** $R^2$ and RMSE values of the chemometric models for norfloxacin in different integration times.

| | | 0.1 s | | 0.2 s | | 0.5 s | |
|---|---|---|---|---|---|---|---|
| **Model** | **Methods** | $R^2$ | **RMSE** | $R^2$ | **RMSE** | $R^2$ | **RMSE** |
| | Raw | 0.7286 | 0.0939 | 0.7606 | 0.0733 | 0.8731 | 0.0476 |
| PLS | WT | 0.8503 | 0.0630 | 0.8747 | 0.0627 | 0.9610 | 0.0446 |
| | LRE | 0.9496 | 0.0296 | 0.9626 | 0.0236 | 0.9784 | 0.0229 |
| | Raw | 0.7803 | 0.0959 | 0.8116 | 0.0894 | 0.9136 | 0.0781 |
| SVM | WT | 0.8673 | 0.0976 | 0.8987 | 0.0789 | 0.9251 | 0.0668 |
| | LRE | 0.9588 | 0.0449 | 0.9665 | 0.0229 | 0.9764 | 0.0210 |

**TABLE 4 |** $R^2$ and RMSE values of the chemometric models for methanol in different integration times.

| | | 0.1 s | | 0.2 s | | 0.5 s | |
|---|---|---|---|---|---|---|---|
| **Model** | **Methods** | $R^2$ | **RMSE** | $R^2$ | **RMSE** | $R^2$ | **RMSE** |
| | Raw | 0.7078 | 1.9980 | 0.8086 | 1.4655 | 0.8458 | 1.3075 |
| PLS | WT | 0.8311 | 0.6551 | 0.8776 | 0.5750 | 0.9178 | 0.4553 |
| | LRE | 0.9017 | 0.5794 | 0.9301 | 0.4692 | 0.9401 | 0.4117 |
| | Raw | 0.7158 | 0.8631 | 0.8382 | 0.7669 | 0.8813 | 0.6148 |
| SVM | WT | 0.8361 | 0.7030 | 0.8701 | 0.6204 | 0.9428 | 0.4506 |
| | LRE | 0.9277 | 0.6417 | 0.9628 | 0.5112 | 0.9768 | 0.3964 |

performance than the conventional WT method in enhancing the prediction accuracy for pharmaceutical quantitative analysis.

As shown in **Table 2**, the raw Raman spectra are all collected in an integration time of 0.2 s. The LRE method is significantly better than the conventional WT method in terms of $R^2$ and RMSE for all components. Quantitation limit (QL) for each pharmaceutical substance is calculated. By definition in ICH guideline (ICH Harmonised Tripartite Guideline, 2005), QL is the lowest concentration of an analyte that can be quantitatively determined with suitable precision and accuracy. It is most often determined as 10 times the standard deviation of the noise from the blank. The LRE method can be used reliably with more than a 15-fold improvement of the practicalQL. Through being processed by the LRE method, QL values for norfloxacin, penicillin potassium, and sulfamerazine are 0.17, 0.13, and 0.19%, respectively. These results reveal that the LRE method can simultaneously improve the performance of quantitative analysis for pharmaceutical multi-component mixtures.

**Table 3** lists $R^2$ and RMSE values of the chemometric models used for quantitative analysis of norfloxacin in different

integration times. The integration times of raw Raman spectra are 0.1, 0.2, and 0.5 s. Raman spectrum's SNR is always proportional to integration time. For evaluating spectral quality, the SNR is defined as the ratio of the peak value of the signal to the root mean square of the noise. For integration times of 0.1, 0.2, and 0.5 s, the average SNR of Raman spectra are 2.47, 3.66, and 6.21, respectively. $R^2$ and RMSE values of the chemometric models for methanol in different integration times are listed **Table 4**. The average SNR of the Raman spectra in the integration times of 0.1, 0.2, and 0.5 s are 2.13, 3.34, and 5.89, respectively.

As shown in **Tables 3**, **4**, the accuracy of the quantitative analysis raises with increasing SNR. According to $R^2$ and RMSE values, it can be proved that the LRE method has a better performance than the conventional WT method. The degree of improvement is higher for low-SNR Raman spectra, which indicates that the LRE method has good noise immunity.

In summary, all predicted results of the Raman spectra preprocessed by the LRE method are in good agreement with corresponding actual values. This method can be applied to improve the accuracy of quantitative analysis based on both PLS and SVM models. It is unrelated to the selection of chemometric models. The LRE method is not restricted by the state of a sample, meaning that it is applicable to both solid and liquid samples. Therefore, it can be regarded as an efficient tool with satisfactory prediction accuracy for pharmaceutical quantitative analysis, especially in the case of low-SNR spectra.

## CONCLUSION

The LRE method has been successfully applied in Raman spectroscopy for pharmaceutical quantitative analysis. It is a simply and feasibly method that can improve the accuracy and robustness of PLS and SVM chemometric models. Our data show that the LRE method has advantages in improving $R^2$ and RMSE for quantitative analysis of pharmaceutical multi-component mixtures, especially in the case of low-SNR spectra. The LRE method will promote the development of Raman spectroscopy in biomedical and pharmaceutical quantitative analysis.

## AUTHOR CONTRIBUTIONS

XM participated in the lab work, supervising lab work, interpretation of data, drafting the manuscript, performing the statistical analysis. XS participated in the lab work, interpretation of data, drafting the manuscript, performing the statistical analysis. HW design of the work, interpretation of data. YW supervised the research, performing the statistical analysis. DC supervised the research, final approval of the version to be published. QL participated in the lab work, supervising lab work, final approval of the version to be published.

## FUNDING

National Key Research and Development Program of China (2017YFC0803603).

## REFERENCES


Ai, Y. J., Liang, P., Wu, Y. X., Dong, Q. M., Li, J. B., Bai, Y., et al. (2018). Rapid qualitative and quantitative determination of food colorants by both Raman spectra and Surface-enhanced Raman Scattering (SERS). *Food Chem.* 241, 427–433. doi: 10.1016/j.foodchem.2017.09.019

Chen, D., Chen, Z., and Grant, E. (2011). Adaptive wavelet transform suppresses background and noise for quantitative analysis by Raman spectrometry. *Anal. Bioanal. Chem.* 400, 625–634. doi: 10.1007/s00216-011-4761-5

Chen, S., Lin, X., Yuen, C., Padmanabhan, S., Beuerman, R. W., and Liu, Q. (2014). Recovery of Raman spectra with low signal-to-noise ratio using Wiener estimation. *Opt. Express* 22, 12102–12114. doi: 10.1364/OE.22.012102

Clupek, M., Matejka, P., and Volka, K. (2007). Noise reduction in Raman spectra: finite impulse response filtrationversusSavitzky-Golay smoothing. *J. Raman Spectrosc.* 38, 1174–1179. doi: 10.1002/jrs.1747

Du, P., Kibbe, W. A., and Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 22, 2059–2065. doi: 10.1093/bioinformatics/btl355

Golbabaee, M., and Vandergheynst, P. (2012). "Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Kyoto), 2741–2744. doi: 10.1109/ICASSP.2012.6288484

Graf, D., Molitor, F., Ensslin, K., Stampfer, C., Jungen, A., Hierold, C., et al. (2007). Spatially resolved Raman spectroscopy of single- and few-layer graphene. *Nano Lett.* 7, 238–242. doi: 10.1021/nl061702a

Guo, X., Yao, Q., and Kwok, J. T. (2017). "Efficient sparse low-rank tensor completion using the Frank-Wolfe algorithm," in *The AAAI Conference on Artificial Intelligence* (San Francisco, CA).

Halko, N., Martinsson, P. G., and Tropp, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* 53, 217–288. doi: 10.1137/090771806

Han, X., Huang, Z.-X., Chen, X.-D., Li, Q.-F., Xu, K.-X., and Chen, D. (2017). On-line multi-component analysis of gases for mud logging industry using data driven Raman spectroscopy. *Fuel* 207, 146–153. doi: 10.1016/j.fuel.2017.06.045

ICH Harmonised Tripartite Guideline (2005). "Validation of analytical procedures: Text and methodology Q2(R1), " in *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use* (Yokohama).

Iordache, M.-D., Bioucas-Dias, J. M., and Plaza, A. (2011). Sparse unmixing of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.* 49, 2014–2039. doi: 10.1109/TGRS.2010.2098413

Jaggi, M. (2013). "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *ICML 2013 - Proceedings of the International Conference on Machine Learning*, Vol. 28. (Atlanta, GA), 427–435.

Johansson, J., Claybourn, M., and Folestad, S. (2010). *Raman Spectroscopy: A Strategic Tool in the Process Analytical Technology Toolbox.* Berlin; Heidelberg: Springer, 241–262.

Kroonenberg, P. M., and Leeuw, J. D. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* 45, 69–97. doi: 10.1007/BF02293599

Li, G. (2008). "Noise removal of Raman spectra using interval thresholding method," in *2008 Second International Symposium on Intelligent Information Technology Application* (Shanghai), 535–539. doi: 10.1109/iita.2008.573

Li, Q., Ma, X., Wang, H., Wang, Y., Zheng, X., and Chen, D. (2018). Speeding up Raman spectral imaging by the three-dimensional low rank estimation method. *Opt. Express* 26, 525–530. doi: 10.1364/OE.26.000525

Li, S., Nyagilo, J. O., Dave, D. P., and Gao, J. X. (2013). Continuous wavelet transform based partial least squares regression for quantitative analysis of Raman spectrum. *IEEE Trans. Nanobiosci.* 12, 214–221. doi: 10.1109/TNB.2013.2278288

Ma, X., Wang, H., Wang, Y., Chen, D., Chen, W., and Li, Q. (2017). Improving the resolution and the throughput of spectrometers by a digital projection slit. *Opt. Express* 25, 23045–23050. doi: 10.1364/OE.25.023045

Neugebauer, U., Clement, J. H., Bocklitz, T., Krafft, C., and Popp, J. (2010). Identification and differentiation of single cells from peripheral blood by Raman spectroscopic imaging. *J. Biophotonics* 3, 579–587. doi: 10.1002/jbio.201000020

Numata, Y., and Tanaka, H. (2011). Quantitative analysis of quercetin using Raman spectroscopy. *Food Chem.* 126, 751–755. doi: 10.1016/j.foodchem.2010.11.059

Pelletier, M. (2003). Quantitative analysis using Raman spectrometry. *Appl. Spectrosc.* 57, 20A–42A. doi: 10.1366/000370203321165133

Ryu, S.-K., Zhao, Q., Hecker, M., Son, H.-Y., Byun, K.-Y., Im, J., et al. (2012). Micro-Raman spectroscopy and analysis of near-surface stresses in silicon around through-silicon vias for three-dimensional interconnects. *J. Appl. Phys.* 111, 063513. doi: 10.1063/1.3696980

Strachan, C. J., Pratiwi, D., Gordon, K. C., and Rades, T. (2004). Quantitative analysis of polymorphic mixtures of carbamazepine by Raman spectroscopy and principal components analysis. *J. Raman Spectrosc.* 35, 347–352. doi: 10.1002/jrs.1140

Tan, Z., Lou, T. T., Huang, Z. X., Zong, J., Xu, K. X., Li, Q. F., et al. (2017). Single-drop raman imaging exposes the trace contaminants in milk. *J. Agric. Food Chem.* 65, 6274–6281. doi: 10.1021/acs.jafc.7b01814

Vickers, T. J., Wambles, R. E., and Mann, C. K. (2001). Curve fitting and linearity: data processing in Raman spectroscopy. *Appl. Spectrosc.* 55, 389–393. doi: 10.1366/0003702011952127

Virtanen, T., Reinikainen, S.-P., Kögler, M., Mänttäri, M., Viitala, T., and Kallioinen, M. (2017). Real-time fouling monitoring with Raman spectroscopy. *J. Memb. Sci.* 525, 312–319. doi: 10.1016/j.memsci.2016.12.005

Yi, C., Lv, Y., Xiao, H., and Tu, S. (2017). Laser induced breakdown spectroscopy for quantitative analysis based on low-rank matrix approximations. *J. Anal. At. Spectrom.* 32, 2164–2172. doi: 10.1039/c7ja 00178a

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 10 October 2018
doi: 10.3389/fchem.2018.00388

# Accuracy Improvement of In-line Near-Infrared Spectroscopic Moisture Monitoring in a Fluidized Bed Drying Process

*Andrey Bogomolov*[1,2]*, *Joachim Mannhardt*[1] and *Oliver Heinzerling*[3]

[1] Blue Ocean Nova GmbH, Aalen, Germany, [2] Samara State Technical University, Samara, Russia, [3] Drug Product Development, AbbVie Deutschland GmbH & Co. KG, Ludwigshafen am Rhein, Germany

An exploratory analysis of a large representative dataset obtained in a fluidized bed drying process of a pharmaceutical powder has revealed a significant correlation of spectral intensity with granulate humidity in the whole studied range of 1091.8–2106.5 nm. This effect was explained by the dependence of powder refractive properties, and hence light penetration depth, on the water content. The phenomenon exhibited a close spectral similarity to the well-known stochastic variation of spectral intensities caused by the process turbulence (the so-called "scatter effect"). Therefore, any traditional scatter-corrective preprocessing incidentally eliminates moisture-correlated variance from the data. To preserve this additional information for a more precise moisture calibration, a time-domain averaging of spectral variables has been suggested. Its application resulted in a distinct improvement of prediction accuracy, as compared to the scatter-corrected data. Further improvement of the model performance was achieved by the application of a dynamic focusing strategy when adjusting the model to a drying process stage. Probe fouling was shown to have a minor effect on prediction accuracy. The study resulted in a considerable reduction of the root-mean-square error of in-line moisture monitoring to 0.1%, which is close to the reference method's reproducibility and significantly better than previously reported results.

**Keywords: fluidized bed drying, moisture monitoring, NIR spectroscopy, light scatter, scatter correction, lighthouse probe, process analytical technology**

OPEN ACCESS

***Edited by:***
*Federico Marini, Università degli Studi di Roma La Sapienza, Italy*

***Reviewed by:***
*Ludovic Duponchel, Université de Lille, France*
*Huawen Wu, BaySpec, Inc., United States*

***Correspondence:**
*Andrey Bogomolov*
*ab@globalmodelling.com*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *16 April 2018*
***Accepted:*** *10 August 2018*
***Published:*** *10 October 2018*

***Citation:***
*Bogomolov A, Mannhardt J and Heinzerling O (2018) Accuracy Improvement of In-line Near-Infrared Spectroscopic Moisture Monitoring in a Fluidized Bed Drying Process. Front. Chem. 6:388. doi: 10.3389/fchem.2018.00388*

## INTRODUCTION

Fluidized bed drying is a common unit operation routinely performed in the pharmaceutical production of solid dosage forms. In a typical batch granulation process, the drying stage immediately follows either the fluidized bed or high-shear granulation stage. It is often considered as one of the most critical steps for achieving stable product quality, i.e., for obtaining granules with desired properties at their minimal variability. Therefore, a close monitoring of the residual moisture content in the process medium is necessary for any quality assurance system in granulate production.

In modern industrial practice, moisture is commonly analyzed in isolated samples. Karl Fischer titration is a classic water analysis technique that has been widely used for decades. A viable alternative accepted by pharmacopeias is thermogravimetric analysis with a drying balance that determines moisture content in the sample as percentage weight loss on drying (LOD). At present,

both techniques are realized as compact desktop devices enabling the at-line analysis of samples taken from a running process.

For the process type studied here, the at-line analysis of granulate moisture content typically takes 20–30 min, representing a good alternative to off-line laboratory analysis of the final product. However, such operability is insufficient to carry out real-time process control, for example by generating alarms on abnormal process states and performing timely corrections. For the same reason, at-line analysis is hardly suitable for accurately determining the process end-point—the time point at which the product reaches its optimal properties. Therefore, instant in-line monitoring of the moisture content in fluidized bed drying is strongly desired to provide a necessary level of process control and to meet growing quality requirements.

Near-infrared (NIR) spectroscopy is an undoubted favorite among real-time sensor systems for moisture monitoring in the production of solids, specifically, in the drying step (Roggo et al., 2007; Burggraeve et al., 2013; Da Silva et al., 2014). In such systems, the diffuse reflectance spectra of the process material are typically measured through an immersion probe. The key advantages of NIR spectroscopy as an in-line analytical technique include the suitability for measurements in media of highly variable bulk density, nondestructiveness, and the capability to place the probe into an appropriate position within the process space while keeping it connected to a remote spectrometer through a fiber optic cable.

The classic NIR spectroscopic moisture analysis relies on two intensive water absorption bands around 1,440 and 1,930 nm, enabling quantitative determination of the moisture in a wide concentration range. In low-selective NIR spectra, the component bands are essentially overlapped and their quantitative analysis requires the application of multivariate modeling, also known as chemometrics. In particular, the partial least-squares (PLS) regression algorithm (Sjöström et al., 1983) is widely accepted in process chemometrics (Bogomolov, 2011).

Over the last decades, the practical acceptance of NIR spectroscopy for in-line moisture monitoring in fluidized bed processing of powders and solids have been constantly growing. Published works (Frake et al., 1997; Rantanen et al., 2000; Zhou et al., 2003; Green et al., 2005; Nieuwmeyer et al., 2007; Skibsted et al., 2007; Luukkonen et al., 2008; Mantanus et al., 2009; Alcalà et al., 2010; Corredor et al., 2011; Peinado et al., 2011; Burggraeve et al., 2012; Demers et al., 2012; Möltgen et al., 2012; Obregón et al., 2013) have focused on the general feasibility of the analysis or on the investigation of specific experimental or modeling aspects (e.g., important process influences, sampling, control strategy, and model transfer). At the same time, the resulting models are typically built and validated on relatively small sets of samples and batches, which can be accounted for by the technical complexity of industrial experiments. Hence, the accuracy estimates reported for similar process setups and conditions are very diverse (Zhou et al., 2003; Green et al., 2005; Nieuwmeyer et al., 2007; Skibsted et al., 2007; Mantanus et al., 2009; Alcalà et al., 2010; Corredor et al., 2011; Peinado et al., 2011; Burggraeve et al., 2012; Demers et al., 2012; Möltgen et al., 2012) and the "ultimate" moisture determination accuracy by in-line NIR spectroscopy under widely variable process conditions remains unknown. Therefore, despite significant progress, the method can hardly be regarded as completely established yet.

In-depth considerations of NIR spectroscopic analysis in terms of light propagation in the complex fluidized bed process medium are rare (Rantanen et al., 2000; Luukkonen et al., 2008; Burggraeve et al., 2013). One of the main obstacles complicating the NIR spectroscopic monitoring of fluidized bed drying is related to process turbulence. A highly variable density of the material around the probe, and consequently the quantity of light reaching the detector, causes intensive random fluctuations of the overall intensity of in-line spectra that are often referred to as the "scatter effect." The problem is commonly resolved by preprocessing the spectra prior to the modeling step. The three most-used scatter correction methods are multiplicative scatter correction (MSC), standard normal variate (SNV), and spectral derivatives (Rinnan et al., 2009). The application of a scatter correction method to in-line process NIR spectra is ubiquitous; no exception has been found in the literature. In most cases, the choice of the preprocessing method is empirical or arbitrary.

In some publications, it was noticed that the NIR spectra expressed in the logarithmic reflectance units (lg(1/R)) exhibited a significant downward shift of the background as the drying progressed (Frake et al., 1997; Rantanen et al., 2000; Zhou et al., 2003; Luukkonen et al., 2008; Burggraeve et al., 2012). Two plausible explanations were suggested, both related to the altering of light scatter conditions in the course of drying. On one hand, the uniform decrease in spectral intensities could be caused by an increase in scattering particle size; this explanation was given by Burggraeve et al. (2012) and Frake et al. (1997). On the other hand, the presence of water on crystal surfaces affects the reflective properties of the granulated powder, resulting in a deeper light penetration and a subsequent higher absorbance of wetter samples (Rantanen et al., 2000; Luukkonen et al., 2008). Rantanen et al. (2000) provided an experimental evidence of the latter phenomenon by using the pharmaceutical excipient (microcrystalline cellulose) as well as inorganic glass beads ("ballotini") with a known size distribution.

The present work aims at building an accurate and robust functional prediction model for in-line moisture content monitoring in fluidized bed drying based on a large representative set of designed process data. Both experimental and modeling factors have been scrutinized to improve the performance of the prediction model. A thorough exploratory data analysis has been applied to help understand the process multivariate trajectory delivered by in-line diffuse-reflectance NIR spectroscopy better. In this study, we focus on efficiently using of the whole spectral information, including both absorption and scatter-related effects of water, to improve the performance of in-line moisture monitoring.

## MATERIALS AND METHODS

Twenty-five pilot-scale fluidized bed drying batches of a pharmaceutical powder mixture were studied by using a 256-pixel diode-array TIDAS 1121 SSG NIR spectrophotometer with

a wavelength range of 1091.8–2106.5 nm (J&M Analytik AG, Germany) that was equipped with the Lighthouse Probe$^{TM}$ (LHP) from GEA Pharma Systems nv – Collette, Belgium (Engler et al., 2009) immersed into the process medium. The LHP was periodically cleaned and recalibrated without process interruption (see section S1.4 of **Supplementary Material**). The total number of cleaning cycles in all batches was 19.

The data of each batch included from 396 to 1,213 NIR spectra collected at 5-s intervals (16,303 spectra in total). In the course of the process, 301 samples of about 5 g (between 5 and 26 samples from each batch) were isolated and analyzed for moisture content as weight loss on drying using a HR73 halogen moisture analyzer (Mettler Toledo GmbH, Switzerland). Reproducibility checks for three LOD analyzers performed during the whole study showed that the measurement standard deviation error does not exceed 0.06% (section S1.2 of **Supplementary Material**).

The main process and the sample information are summarized in **Table S-1**. Out of the 301 samples, three were rejected from further analysis as evident outliers (section S2.3.1 of **Supplementary Material**).

Individual batch conditions were set in accordance with a developed experimental design to cover the whole range of practical process variability. Moisture content in the selected samples varied between 2.38 and 25.92%. The active pharmaceutical ingredient (API) was present in four assay levels: 0 (placebo), 0.1, 1.0, and 10.0 mg. The range of process temperatures was 30.5–49.7°C. Eight batches (88 reference samples) formed a validation subset that was representative of the process conditions and used for model validation; the other 17 batches were used as the calibration set in that case (**Table S-1**).

A subset of 101 experimental samples were additionally analyzed off-line by using an MPA Fourier-transform (FT-) IR spectrometer (Bruker, Germany) with an integrating sphere (section S1.5 of **Supplementary Material**).

Principal component analysis (PCA) and PLS regression are multivariate data analysis algorithms described in the literature (Sjöström et al., 1983; Wold et al., 1987). The multivariate spaces, namely, PCA model principal components (PCs) and PLS latent variables (LVs) represented by their score (**t**) and loading (**p**) vectors, were used for exploratory data analysis. Conventional data preprocessing methods employed were MSC, SNV, and first-derivative using the Savitzky–Golay smoothing filter, as described by Rinnan et al. (2009).

Three validation techniques were applied with each regression model: leave-one(-sample)-out (LOO), a.k.a. full cross-validation (CV), leave-a-batch-out (LBO) CV, and validation by a preselected set (**Table S-1**). The performance of the models was characterized by root-mean-square errors (*RMSE*) of calibration, validation, and prediction, as well as corresponding determination coefficients $R^2$.

A detailed description of data acquisition and analysis is given in section S1 of **Supplementary Material**.

# RESULTS AND DISCUSSION

## Exploratory Analysis of In-line Spectral Data

**Figure 1** presents a set of 1,213 in-line NIR spectra obtained in batch B03 (**Table S-1**). An expected intensity reduction of the main water band in the 1,920–1,940 nm range during the process is clearly observed. Another distinct feature is the high variability of spectral intensities over the whole wavelength range (the so-called "scatter effect"), caused by strong instant density fluctuations of the granulate (and its spatial distribution) around the probe.

At the same time, the overall spectral intensity tends to fall gradually during the process, generally following the dynamics of water reduction. This trend can be illustrated by the time dependencies of the spectral intensity at two separate wavelengths: 1932.0 nm at the maximum of the main water band and 1708.1 nm where no noticeable water absorption is expected. Both intensities strongly correlate with the reference moisture content (**Figure 2A**). Data smoothing along the time scale makes this correlation even more distinct.

The moisture- and time-dependent changes in the batch processes can be effectively visualized by using data animation (section S2.1 and **Video S-1**, **Supplementary Material**). Animated spectral data reveal the same trends, namely water band reduction and stochastic background variation accompanied by a gradual fall of the spectrum intensity in the whole range.

In this situation, preprocessing is desirable, but it should be applied to the data variable vectors, i.e., along the time scale, as shown in **Figure 2A**. As the turbulence effect is supposed to be pure noise, the smoothing of variables is a straightforward way to eliminate it with a minimal loss of the informative variance.

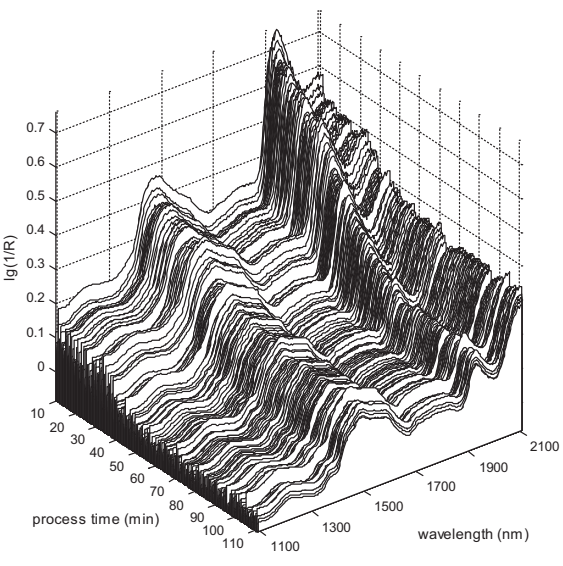



**FIGURE 1 |** In-line NIR spectra in batch B03.

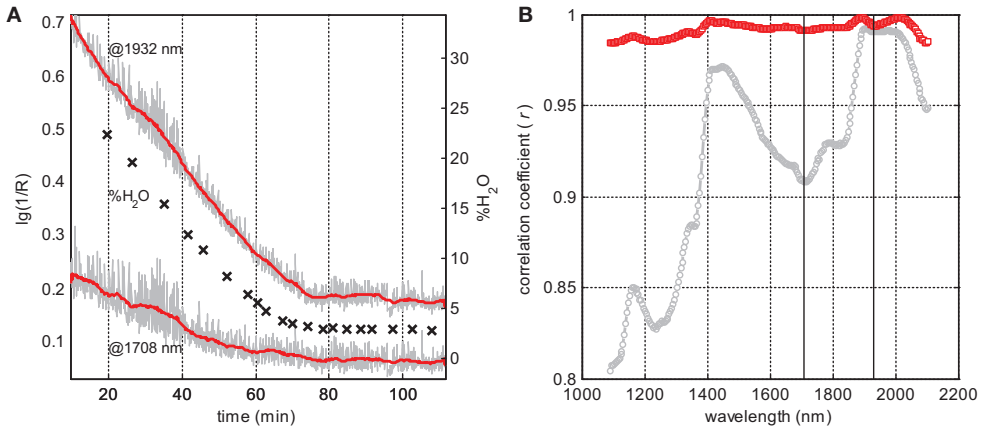



**FIGURE 2 |** Exploratory analysis of B03 data: **(A)** raw (gray line) and smoothed (red line) spectral intensities at the two selected wavelengths and reference moisture content (crosses) vs. process time; and **(B)** correlation coefficients between the moisture content and spectral intensities at individual wavelengths for raw (gray circles) and smoothed (red squares) data; vertical lines at 1,708 and 1,932 nm correspond to the dependencies presented in **(A)**; data were smoothed with a 47-point window.

One of the simplest smoothing techniques, the moving window averaging algorithm, has been used to preprocess the matrix of spectral data **X**. In this method, each element $x_{ij}$ in **X,** where $i$ and $j$ are respectively the object (spectrum) and variable (wavelength) indices, is replaced by a corrected value $x^s_{ij}$ calculated as a mean of the surrounding points within a window having the width defined by an odd number $k$ (Equation 1):

$$x^s_{ij} = \frac{\sum_{i=i-(k-1)/2}^{i=i+(k-1)/2} x_{ij}}{k} \quad (1)$$

The transformation is performed for each variable in **X**. $(k - 1)/2$ end-points on each side of the variable vector were smoothed with a reduced window of $(l - 1) \cdot 2 + 1$ points, where $l$ is the point ordinal number from either spectrum end.

Data averaging within a selected time window is similar to a respective enhancement of the spectrum acquisition time, thus enlarging the virtual sample size captured by a single measurement. However, in contrast to the measurement time adjustment, the mathematical averaging does not place any limit on the time step of data acquisition, i.e., it can be performed with a time window that is much wider than the physical step size. A positive effect of the variable smoothing for the modeling of a fermentation process data has been reported (Skibsted et al., 2001).

Pair-wise correlations between the LOD values and the intensities at individual variables in the corresponding (closest to the sampling times) in-line spectra were analyzed in the whole wavelength range. **Figure 2B** presents linear correlation coefficients ($r$) as a function of wavelength in B03. All spectral variables exhibit a strong intensity correlation with the moisture content, even in the raw data. Eliminating the process noise using the suggested averaging method (Equation 1) results in a dramatic enhancement of $r$. It also looks natural that correlation maxima are observed around major water bands. However, even beyond the water absorbance regions, this correlation is very high. Thus, the lowest $r$ observed in B03 at the short-wave end of the spectral region is still greater than 0.8 (**Figure 2B**); after the smoothing, this value increases to 0.98. Similar dependencies were observed for all the 25 studied batches.

A high correlation of lg(1/R) with the moisture content in the whole studied NIR range is in agreement with some published observations. This phenomenon can be explained by altering the refractive properties of the granulate (Rantanen et al., 2000). Indeed, in the course of drying, the liquid bridges holding the primary particles together (Burggraeve et al., 2013) are replaced by air. The crystal–air interface is characterized by a higher difference of refractive indices than the crystal–water pair. Thus, drying leads to a higher scatter—and hence an increased quantity of diffusely reflected light reaching the detector—that corresponds to a decrease in the spectral intensity expressed in absorbance type of units. For relatively large particles constituting the granules, this effect should be wavelength-independent. An intuitive illustration of the particle wetting effect and its uncomplicated explanation using the representative layer theory was given by Dahm (2013). A similar correlation of the Raman spectral background with the moisture content was observed in our earlier studies on pellet coating (Bogomolov et al., 2010) and granulation process monitoring (Bogomolov, 2011), and was also explained by the effect of moisture on the light propagation conditions in the process medium. Considering the strength of the spectrum variable correlation with the moisture content observed in the whole range of process conditions studied, an earlier explanation of the phenomenon in terms of changing particle size distribution during the drying course (Frake et al., 1997; Burggraeve et al., 2012) has not been confirmed. This hypothesis does not agree with the complex shape of the correlation curve in **Figure 2B**. Particle size distribution can be a minor water-correlated factor affecting the spectra of the drying process, though.

The effect of humidity on the light penetration depth in porous materials can be compared to the watermark technique commonly used for banknote authentication. The very name of watermarks comes from the visual similarity of paper thickness variation and its wetting effects, both resulting in a decrease in the back-scattered light. Darkening of wetted powders (e.g., sand) is another manifestation of the same phenomenon that is not limited to the visible light and should be inherent in any material with a highly developed surface. The spectral variance related to the changing refractive properties of the powder is also expected to be present in the in-line process spectra. However, being wavelength-independent, the moisture-related spectral changes are masked by the stochastic "scatter effect" and then eliminated by any scatter correction. Earlier studies on in-line moisture analysis by using NIR spectroscopy neither paid any significant attention to the analytical information hidden in the "watermarks" nor attempted to use it in the modeling.

A deeper insight into the data structure and its modification by adopting different preprocessing methods was obtained by the PCA of augmented process data (section S2.2 of **Supplementary Material**) that makes possible the investigation of process trajectories of individual batches in the same multivariate factor space.

As one can see from the scores of batch B10 taken as an example here (**Figure 3** and **Figure S-3**), the first PC (95.49% of **X**-variance) of the raw-data model (**Figure 3A**) is strongly associated with the moisture content, while $PC_2$ (4.23%) basically describes the process turbulence. A remarkable similarity of the first two loadings (**Figure S-4a**) with the correlation coefficient $r = 0.998$ is a confirmation of a close spectral affinity of these two phenomena. A scatter-driven correlation of spectral intensities with the moisture content is confirmed by the uniformly positive $\mathbf{p}_1$. A simultaneous presence of the water absorption peaks in this plot implies that $PC_1$ tends to capture the whole variance due to the moisture reduction, related to both absorbance and scatter phenomena.

Although the process noise is basically described by $PC_2$, it strongly pollutes $PC_1$ and all further components in the raw-data model. The suggested smoothing method effectively eliminates this noise from the model scores (**Figures 3B,C** and **Figures S-3b,c**) without any essential change to the loadings (**Figures S-4b,c**). In contrast, the SNV, MSC, and first derivative (**Figures S-4d-f**)) strongly modify the whole factor space; they essentially remove random fluctuations from the first two score vectors (less noisy for the first derivative) but further PCs stay very noisy (**Figures 3D–F**). The smoothed data is suitable for exploring the process trajectories in the PCA factor space. Most of the minor features revealed in the refined scores $\mathbf{t}_2$-$\mathbf{t}_7$ (**Figures 3B–F**) can be assigned to certain process events, i.e., to changing process phases or LHP cleaning cycles. The PCA score plots for all batches can be found in **Figure S-3**.

**X**-variances captured by individual PCs (**Table S-3** in section S2.2.2 of **Supplementary Material**) indicate at least six significant factors for all preprocessing methods, while the $PC_8$–$PC_{10}$ are definitely negligible. The $PC_7$ seems to be a boundary case, and its significance should be proved by using other criteria. Considering spectrum-like loadings (**Figure S-4**) and process-reflecting scores, in particular in the time-wise averaged data (**Figure 3C**), seven PCs are likely to be relevant. Additional considerations helping to deduce a number of PCs in the augmented process data are considered in section S2.2 of **Supplementary Material**.

In general, the low variances captured by minor principal components $PC_2$-$PC_7$ (**Table S-3**) illustrate a much higher sensitivity of NIR spectroscopy to water than to other chemical or physical variability sources in the drying process medium. Nevertheless, a thorough study of the complete PCA model resulted in some practically important observations. Thus, LHP fouling and cleaning during the process has a minor effect on the observed in-line spectra, in particular, at the final process stage (section S2.2.1 of **Supplementary Material**).

An exploratory data analysis performed has revealed an essential correlation of all spectral variables with the moisture content. The PCA analysis of the united dataset (16,303 spectra) has shown that this effect is overlaid with a variation on the stochastic spectrum intensity caused by the process noise. Since both scatter-driven effects have similar spectral signatures, the application of conventional normalization or derivative preprocessing methods of scatter correction incidentally removes useful information contained in the spectrum background. Instead, it was suggested to perform the smoothing of spectral variables along the time domain, e.g., using a moving window average.

## Building an Accurate PLS Regression Model of Moisture Content

For efficiently using the additional moisture-related information contained in the spectral variables, the dependence of model accuracy on averaging window width (WW = $k$ points) has been studied. PLS models for all possible odd $k$ values between 3 and 101 in different moisture ranges were compared (section S2.3.2 in **Supplementary Material**). Since the in-line smoothing of time dependencies results in a delay of $2k - 1$ trajectory points (half WW) between the process and analysis times (Bogomolov, 2011), light smoothing is technically preferred. WW = 15 was found to be optimal in all cases as it provided an essential improvement of the model accuracy with a reasonable delay of 35 s. The full WW of 70 s approximately corresponds to a material circulation period in this process and dryer type. Thus, each portion of the granulate has a good chance of being exposed to spectroscopic measurement during this time. Due to the averaging, a virtual sample size captured by spectroscopic measurement, and hence the level of scrutiny of analysis, is extended. From this point of view, an optimal WW should correspond to an averaged spectrum that is representative of the bulk material volume, while remaining a nearly instant measurement compared to the total process time. This principle can be suggested as a rule of thumb for optimal data averaging in the drying process analysis and similar applications. A 47-point averaging was found to be a "global" optimum in our case; stronger smoothing does not lead to any significant gain. Based on these observations, 15- and 47-point smoothing windows have been chosen as benchmarks for model comparison (the respective preprocessing methods are designated as S15 and S47). **Table 1** presents a summary

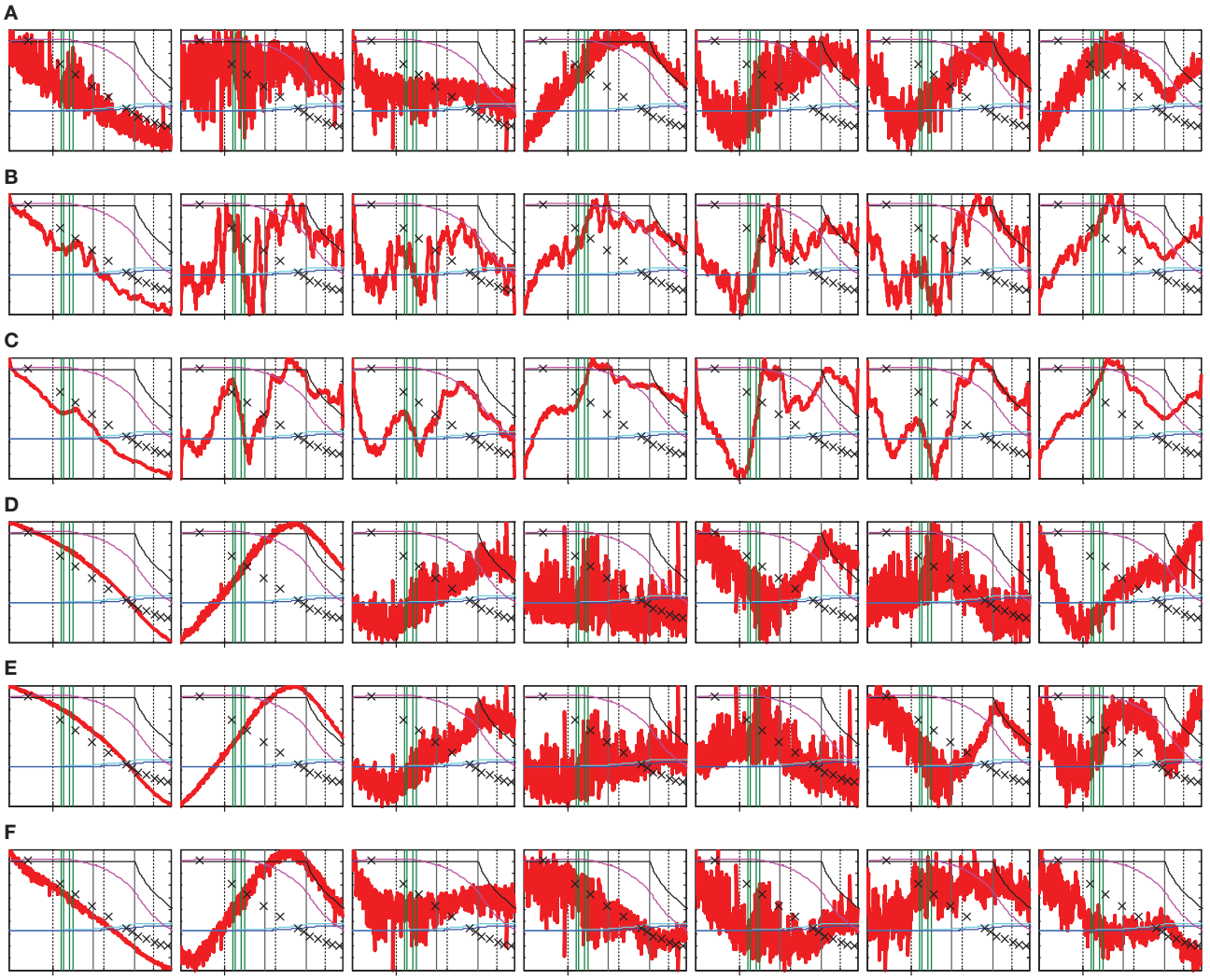



**FIGURE 3 |** PCA scores (vertical axis, arbitrary units) vs. process time (horizontal axis, process time from 1,130 to 4,331 s, with the tick at 2,000 s) for batch B10. The plots in a line present individual scores $\mathbf{t}_1$-$\mathbf{t}_7$ (left to right) for different data preprocessing methods: **(A)** none; **(B,C)** variable smoothing with 15- and 47-point windows, respectively; **(D)** MSC; **(E)** SNV; and **(F)** first derivative using the Savitzky–Golay smoothing filter. Process parameters are shown overlaid: moisture content in reference samples (crosses); drying air temperature (black line), product and exhaust air temperatures (light and dark blue lines, respectively); exhaust air humidity (violet line); and LHP cleaning start/end points (vertical green lines).

of full-spectrum modeling results for different moisture ranges, preprocessing techniques, and validation methods.

The data covers a wide range of moisture contents from 2 to 26% (**Table S-1**). As the prediction error may be nonuniform depending on the drying stages (Mantanus et al., 2009), several PLS models were built corresponding to moisture LOD ranges <20% ($D_{20}$), <15% ($D_{15}$), and <10% ($D_{10}$), in addition to the full-data (D) models. The abundance of measurement points makes possible the use of this data reduction without a significant impact to the model quality. The upper value of moisture content noticeably reduces the *RMSE* (e.g., for LBO CV, it falls from 0.21 in D to 0.13 in $D_{10}$), keeping $R^2$ at the same high level of 0.997–0.998 (**Table 1**). A strong error dependence on the moisture content can be practically employed to improve the performance of moisture monitoring in general. Thus, prediction software can switch to a more precise model as soon as a certain moisture content level is reached, providing an automatic model "focusing" in the process course. By this way, the most critical final stage of drying can be monitored with the highest accuracy.

A number of LVs to be kept in PLS models was estimated from the *RMSE* of different validation methods and from the explained **X**- and **y**-variances (**Table S-4**). **Figure 4** compares the LBO CV *RMSE* dependencies on the number of LVs for the models in different moisture ranges (**Figure 4A**) and data averaging degrees (**Figure 4B**). Their common trend is that the validation error reaches a plateau starting from the seventh LV; faint minima at higher factor numbers do not seem significant. Note that LBO

**TABLE 1 |** PLS regression statistics for in-line moisture content determination: model comparison for different moisture ranges and preprocessing techniques using different validation methods; all models were built with 7 LVs.

| Data[a] | nS[b] | PP[c] | Calibration[d] | | LOO CV[e] | | LBO CV[f] | | Validation set[g] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | *RMSE* | $R^2$ | *RMSE* | $R^2$ | *RMSE* | $R^2$ | *RMSE* | $R^2$ |
| D | 298 | None | 0.207 | 0.9981 | 0.222 | 0.9979 | 0.236 | 0.9976 | 0.246 | 0.9977 |
| | | S15[h] | 0.181 | 0.9986 | 0.194 | 0.9984 | 0.209 | 0.9981 | 0.198 | 0.9984 |
| | | S47[i] | 0.178 | 0.9986 | 0.191 | 0.9984 | 0.210 | 0.9981 | 0.197 | 0.9984 |
| | | MSC | 0.264 | 0.9970 | 0.292 | 0.9963 | 0.341 | 0.9950 | 0.290 | 0.9969 |
| | | SNV | 0.312 | 0.9958 | 0.342 | 0.9949 | 0.395 | 0.9933 | 0.334 | 0.9961 |
| | | 1D2.15[j] | 0.203 | 0.9982 | 0.221 | 0.9979 | 0.251 | 0.9973 | 0.256 | 0.9977 |
| $D_{20}$ | 289 | None | 0.190 | 0.9979 | 0.205 | 0.9976 | 0.216 | 0.9973 | 0.269 | 0.9967 |
| | | S15 | 0.169 | 0.9983 | 0.182 | 0.9981 | 0.195 | 0.9978 | 0.208 | 0.9979 |
| | | S47 | 0.166 | 0.9984 | 0.178 | 0.9982 | 0.190 | 0.9979 | 0.211 | 0.9979 |
| $D_{15}$ | 268 | None | 0.152 | 0.9978 | 0.161 | 0.9975 | 0.170 | 0.9973 | 0.146 | 0.9979 |
| | | S15 | 0.146 | 0.9980 | 0.155 | 0.9977 | 0.163 | 0.9975 | 0.137 | 0.9981 |
| | | S47 | 0.139 | 0.9982 | 0.147 | 0.9979 | 0.155 | 0.9977 | 0.129 | 0.9984 |
| | | MSC | 0.175 | 0.9971 | 0.188 | 0.9967 | 0.210 | 0.9958 | 0.209 | 0.9963 |
| | | SNV | 0.175 | 0.9971 | 0.191 | 0.9965 | 0.215 | 0.9957 | 0.184 | 0.9970 |
| | | 1D2.15 | 0.153 | 0.9978 | 0.164 | 0.9974 | 0.181 | 0.9969 | 0.158 | 0.9976 |
| $D_{10}$ | 213 | None | 0.116 | 0.9967 | 0.124 | 0.9962 | 0.141 | 0.9951 | 0.129 | 0.9946 |
| | | S15 | 0.109 | 0.9971 | 0.116 | 0.9967 | 0.132 | 0.9957 | 0.122 | 0.9950 |
| | | S47 | 0.109 | 0.9971 | 0.116 | 0.9967 | 0.137 | 0.9954 | 0.121 | 0.9952 |

*[a]Dataset used: D – full dataset, $D_{20}$, $D_{15}$, and $D_{10}$ – datasets limited to LOD moisture content below 20, 15, and 10%, respectively; [b]the number of samples without outliers (see section S2.3 of* ***Supplementary Information****); [c]preprocessing applied; [d]calibration statistics; [e]full cross-validation statistics; [f]leave-a-batch-out cross-validation statistics; [g]validation set (****Table 1****) prediction statistics; [h]variable averaging with 15-point window; [i]variable averaging with 47-point window; [j]Savitzky–Golay first derivative with second-order polynomial and 15-point smoothing window.*

CV is generally the most conservative (i.e., resulting in the highest errors) validation method in **Table 1**. Data scatter correction does not result in any model simplification as expected. **Figure 4B** shows that the validation *RMSE* for MSC-preprocessed $D_{15}$ data is even higher than the *RMSEV* of the model obtained after moderate (S15) data smoothing. This effect is observed for any number of LVs higher than one. Starting from the sixth LV, the prediction error after MSC becomes even worse than in the raw-data model. This behavior agrees with the earlier PCA-based conclusion that conventional scatter correction refines only the two first factors of the multivariate space, transferring the process noise into higher yet significant model dimensions.

The analysis of the captured **X**- and **y**-variances (**Table S-4**) exhibited similar trends. It was also shown that seemingly insignificant variances captured by the seventh LV in the calibration data are still in agreement with the respective precisions of the NIR spectrometer and the LOD analyzer (section S2.3.3 of **Supplementary Material**).

The first two PLS loadings (**Figure S-6**) are almost identical to those in the augmented PCA (**Figure S-4**); therefore, both multivariate modeling spaces are essentially the same. Meaningful shapes of the first seven loadings, which are similar in PCA (**Figure S-4**) and PLS models (**Figure S-6**) as well as PCA scores (**Figure 3**), provide an additional justification of the chosen model's complexity. The noticeable positive offset of $\mathbf{p}_1$ in raw and smoothed data models (**Figures S-6a-c**) indicates that PLS regression makes use of both absorbance and scatter-correlated variances for moisture calibration. The loadings $\mathbf{p}_3$ to $\mathbf{p}_7$ still exhibit similar (as in PCA) interpretable spectrum-like features. Therefore, seven LVs were found to be optimal for all moisture ranges and data preprocessing methods, consistent with the earlier PCA result for all spectral data. This number is also reasonable, considering the physical and chemical complexity of the process as well as the anticipated nonlinearity of spectral responses. It is also acceptable from the point of view of calibration set size.

**Table S-4** also confirms the efficiency of variable smoothing. Cumulative **y**-variances grow with the averaging WW, reducing a misbalance between the **X**- and **y**-variances for any number of LVs, in particular, for $LV_1$. Starting from $LV_3$, the **y**-variance captured in smoothed data becomes higher than that in the models preceded by scatter correction (e.g., MSC). In detail, the problem of deducing the optimal number of LVs is considered in section S2.3.3 (**Supplementary Material**).

Validation statistics presented in **Table 1** evidences that the suggested data averaging approach is advantageous as compared to the MSC, SNV, and first derivative using the Savitzky–Golay smoothing filter. It is remarkable that any scatter correction (most essentially, MSC or SNV) leads to higher calibration and validation errors than those for raw spectral data (This comparison is provided for D and $D_{15}$, but it holds for all datasets). **Figure 5** illustrates the model performance achieved in $D_{15}$.

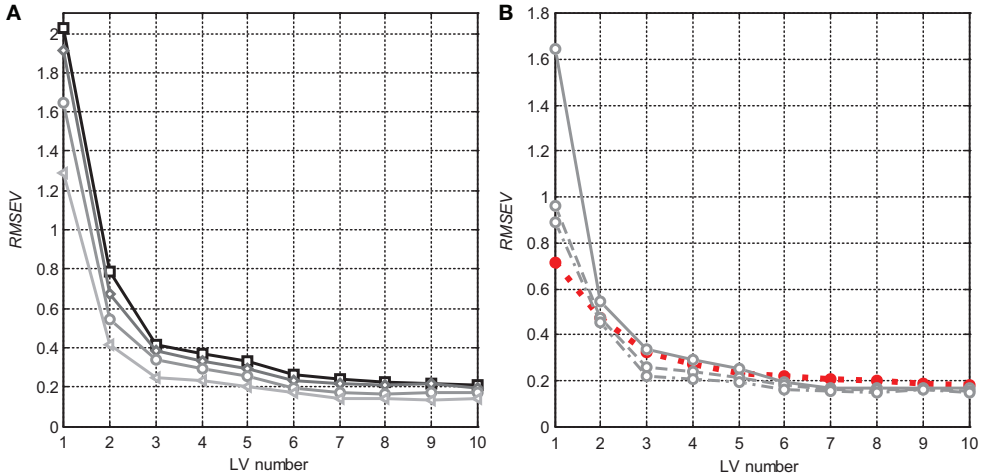



**FIGURE 4 |** *RMSE* dependencies (LBO CV) on the number of LVs in PLS models: **(A)** for nonpreprocessed data in different moisture content ranges: D (squares), $D_{20}$ (diamonds), $D_{15}$ (circles), and $D_{10}$ (triangles); and **(B)** $D_{15}$ data with different smoothing degrees: none (solid), S15 (dashed), and S47 (dash-dotted), as well as for MSC preprocessing (red dotted, filled markers).

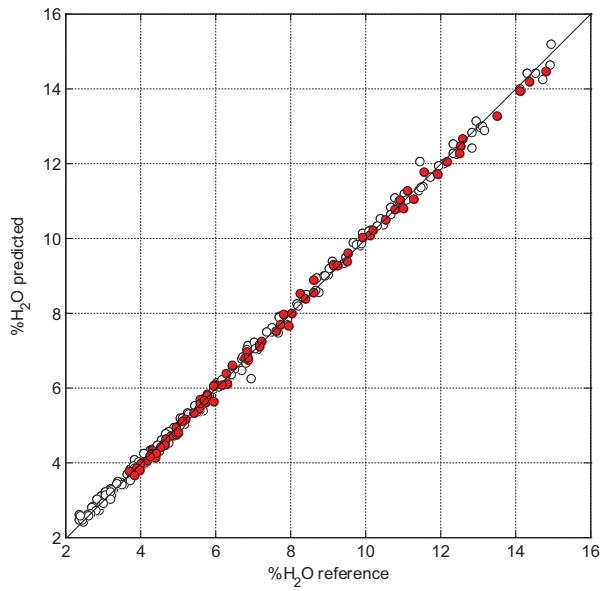



**FIGURE 5 |** PLS predicted (7 LVs) vs. measured moisture content for $D_{15}$ with 15-point smoothing; calibration and validation samples are presented by hollow and red-filled markers, respectively.

A subset of 101 process samples was additionally analyzed off-line by using a high-resolution FT-NIR spectrometer (section S2.3.5 of **Supplementary Material**). The integration sphere applied in this case excluded any scatter-related stochastic variation of spectral intensities. Nevertheless, all spectral variables (including the background signal) exhibited the same strong correlation with the sample moisture content (**Figures S-9, S-10**), as in the case of in-line spectra (**Figure 2**). This fact confirms our previously given explanation of this effect in terms of changing light propagation conditions. Moreover, the performance of the PLS model built on 96 off-line spectra (samples with LOD > 15% were used) was found to be essentially the same (cross-validation *RMSE* = 0.108) as in the model built on respective averaged in-line spectra (S15) of the same process samples (**Table S-6**). This remarkable result provides an additional confirmation of the efficiency of the suggested method. For more details on the off-line analysis results, see **Supplementary Material**, section S2.3.5.

The time dependencies of the predicted moisture content in B12 (**Figure S-7**) illustrate the additional advantages of the suggested preprocessing technique. Variable smoothing most efficiently eliminates the noise contained in process trajectories at the beginning of the drying process, when the moisture content is greater than 15%. It also helps avoid prediction artifacts related to probe cleaning during the "wet" process stage. Section S2.3.4 in **Supplementary Material** provides a detailed discussion of the predicted drying trajectories.

In numerous publications on in-line diffuse-reflectance NIR monitoring of fluidized bed drying and similar processes, data analysis is always prefaced by MSC, SNV, or derivatives without exception. A mandatory application of corrective preprocessing may only be justified in preliminary feasibility studies, when the small calibration/validation dataset does not allow for building models of adequate complexity. The results reported here could be used as evidence for the destructiveness of scatter correction for the moisture calibration, as it eliminates a significant portion of the useful variance. Similar ideas have been formulated in the literature (Chen and Thennadil, 2012), where the information content of MSC coefficients was analyzed. The PLS capability of employing the quantitative information delivered by the scatter has earlier been illustrated in other applications, in particular in particle size analysis (Nieuwmeyer et al., 2007) and the quantitative determination of fat and protein in milk (Bogomolov

et al., 2012). In these cases, the predictive models built on raw data exhibited a noticeably better performance, as compared to those in which any scatter-correction was applied. For in-line process data, the suggested smoothing approach, performed in a time rather than spectral domain, presents a viable alternative to the classic scatter correction of spectra, to eliminate noise while preserving useful information contained in the spectral variables.

## CONCLUSIONS

In light of our presented results, the following recommendations to practical NIR spectroscopic monitoring of moisture content in fluidized bed drying and similar process types can be formulated. A very common practice of *a priori* scatter correction of in-line process spectra prior to the multivariate calibration is generally discouraged, because it may eliminate an essential part of the water-related variance from the data and thus deteriorate the resulting prediction model. To avoid this, quantitative modeling should be prefaced by an exploratory analysis of the raw data to investigate the relevance of both absorbance and scatter-related effects of moisture by using a sufficiently large representative set of designed samples and process conditions. These considerations are equally valid in cases when water content is not directly determined, but it should be taken into account by an accurate multivariate model as an important process factor. Process noise, i.e., stochastic background and intensity variations of in-line spectra, can be efficiently eliminated with a minimal loss of useful information by means of data smoothing along the time scale. The parameters of smoothing strengths should be adjusted depending on the process scale and dynamics. Building accurate quantitative models should rely on a methodically determined number of latent variables. A deliberate application of less LVs than their optimal number following from the model diagnostics—sometimes done by researchers to guarantee an avoidance of overfitting—is not always justified. An underfitting may often be more undesirable for model prediction accuracy.

## AUTHOR CONTRIBUTIONS

AB conceived and wrote the paper and analyzed the data. JM conceived the project and organized and planned industrial experiments. OH performed the experiments and analyzed the data.

## ACKNOWLEDGMENTS

The Ministry of Education and Science of the Russian Federation supported this work within the framework of the basic part of state task on the theme Adaptive technologies of analytical control based on optical sensors (Project No. 4.7001.2017/BP). The authors thank Tomas Vermeire (GEA, Belgium) and the colleagues from the previous Pharmaceutical and Analytical Development department (Weesp, The Netherlands) for their support of the experiments. J&M Analytik AG is acknowledged for organization and support. Prof. Dr. Rudolf W. Kessler (Reutlingen University, Germany) is acknowledged for fruitful discussions. Ivan and Petr Bogomolov helped in manuscript preparation.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00388/full#supplementary-material

## REFERENCES

Alcalà, M., Blanco, M., Bautista, M., and González, J. M. (2010). On-Line monitoring of a granulation process by NIR spectroscopy. *J. Pharm. Sci.* 99, 336–345. doi: 10.1002/jps.21818

Bogomolov, A. (2011). Multivariate process trajectories: capture, resolution and analysis. *Chemom. Intell. Lab. Syst.* 108, 49–63. doi: 10.1016/j.chemolab.2011.02.005

Bogomolov, A., Dietrich, S., Boldrini, B., and Kessler, R. W. (2012). Quantitative determination of fat and total protein in milk based on visible light scatter. *Food Chem.* 134, 412–418. doi: 10.1016/j.foodchem.2012.02.077

Bogomolov, A., Engler, M., Melichar, M., and Wigmore, A. (2010). In-line analysis of a fluid bed pellet coating process using a combination of near infrared and Raman spectroscopy. *J. Chemom.* 24, 544–557. doi: 10.1002/cem.1329

Burggraeve, A., Monteyne, T., Vervaet, C., Remon, J. P., and de Beer, T. (2013). Process analytical tools for monitoring, understanding, and control of pharmaceutical fluidized bed granulation: a review. *Eur. J. Pharm. Biopharm.* 83, 2–15. doi: 10.1016/j.ejpb.2012.09.008

Burggraeve, A., Silva, A. F., van den Kerkhof, T., Hellings, M., Vervaet, C., Remon, J. P., et al. (2012). Development of a fluid bed granulation process control strategy based on real-time process and product measurements. *Talanta* 100, 293–302. doi: 10.1016/j.talanta.2012.07.054

Chen, Y.-C., and Thennadil, S. N. (2012). Insights into information contained in multiplicative scatter correction parameters and the potential for estimating particle size from these parameters. *Anal. Chim. Acta* 746, 37– 46. doi: 10.1016/j.aca.2012.08.006

Corredor, C. C., Both, D., and Bu, D. (2011). Comparison of near infrared and microwave resonance sensors for at-line moisture determination in powders and tablets. *Anal. Chim. Acta* 696, 84–93. doi: 10.1016/j.aca.2011.03.048

Da Silva, C. A. M., Butzge, J. J., Nitz, M., and Taranto, O. P. (2014). Monitoring and control of coating and granulation processes in fluidized beds – A review. *Adv. Powder Technol.* 25, 195–210. doi: 10.1016/j.apt.2013.04.008

Dahm, D. J. (2013). Explaining some light scattering properties of milk using representative layer theory. *J. Near Infrared Spectrosc.* 21, 323–339. doi: 10.1255/jnirs.1071

Demers, A.-M., Gosselin, R., Simard, J.-S., and Abatzoglou, N. (2012). In-line near infrared spectroscopy monitoring of pharmaceutical powder moisture in a fluidized bed dryer: an efficient methodology for chemometric model development. *Can. J. Chem. Eng.* 90, 299–303. doi: 10.1002/cjce.20691

Engler, M., Bogomolov, A., and Mannhardt, J. (2009). Die Lighthouse-Probe, eine neuartige Sonde für die Prozessanalytik. *Chem. Ing. Tech.* 81, 1114–1115. doi: 10.1002/cite.200950354

Frake, P., Greenhalgh, D., Grierson, S. M., Hempenstaal, J. M., and Rudd, D. R. (1997). Process control and end-point determination of a fluid bed granulation by application of near infra-red spectroscopy. *Int. J. Pharm.* 151, 75–80. doi: 10.1016/S0378-5173(97)04894-1

Green, R. L., Thurau, G., Pixley, N. C., Mateos, A., Reed, R. A., and Higgins, J. P. (2005). In-line monitoring of moisture content in fluid bed dryers using near-IR

spectroscopy with consideration of sampling effects on method accuracy. *Anal. Chem.* 77, 4515–4522. doi: 10.1021/ac050272q

Luukkonen, P., Fransson, M., Björn, I. N., Hautala, J., Lagerholm, B., and Folestad, S. (2008). Real-time assessment of granule and tablet properties using in-line data from a high-shear granulation process. *J. Pharm. Sci.* 97, 950–959. doi: 10.1002/jps.20998

Mantanus, J., Ziémons, E., Lebrun, P., Rozet, E., Klinkenberg, R., Streel, B., et al. (2009). Moisture content determination of pharmaceutical pellets by near infrared spectroscopy: method development and validation. *Anal. Chim. Acta* 642, 186–192. doi: 10.1016/j.aca.2008.12.031

Möltgen, C.-V., Puchert, T., Menezes, J. C., Lochmann, D., and Reich, G. (2012). A novel in-line NIR spectroscopy application for the monitoring of tablet film coating in an industrial scale process. *Talanta* 92, 26–37. doi: 10.1016/j.talanta.2011.12.034

Nieuwmeyer, F. J., Damen, M., Gerich, A., Rusmini, F., van der Voort Maarschalk, K., and Vromans, H. (2007). Granule characterization during fluid bed drying by development of a near infrared method to determine water content and median granule size. *Pharm. Res.* 24, 1854–1861. doi: 10.1007/s11095-007-9305-5

Obregón, L., Quiñones, L., and Velázquez, C. (2013). Model predictive control of a fluidized bed dryer with an inline NIR as moisture sensor. *Cont. Eng. Pract.* 21, 509–517. doi: 10.1016/j.conengprac.2012.11.002

Peinado, A., Hammond, J., and Scott, A. (2011). Development, validation and transfer of a near Infrared method to determine in-line the end point of a fluidised drying process for commercial production batches of an approved oral solid dose pharmaceutical product. *J. Pharm. Biomed. Anal.* 54, 13–20. doi: 10.1016/j.jpba.2010.07.036

Rantanen, J., Räsänen, E., Tenhunen, J., Känsäkoski, M., Mannermaa, J.-P., and Yliruusi, J. (2000). In-line moisture measurement during granulation with a four-wavelength near infrared sensor: an evaluation of particle size and binder effects. *Eur. J. Pharm. Biopharm.* 50, 271–276. doi: 10.1016/S0939-6411(00)00096-5

Rinnan, Å., van den Berg, F., and Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Anal. Chem.* 28, 1201–1222. doi: 10.1016/j.trac.2009.07.007

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., and Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *J. Pharm. Biomed.* 44, 683–700. doi: 10.1016/j.jpba.2007.03.023

Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å., and Martens, H. (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Anal. Chim. Acta* 150, 61–70. doi: 10.1117/12.2227906

Skibsted, E., Lindemann, C., Roca, C., and Olsson, L. (2001). On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration. *J. Biotechnol.* 88, 47–57. doi: 10.1016/S0168-1656(01)00257-7

Skibsted, E. T., Westerhuis, J. A., Smilde, A. K., and Witte, D. T. (2007). Examples of NIR based real time release in tablet manufacturing. *J. Pharm. Biomed. Anal.* 43, 1297–1305. doi: 10.1016/j.jpba.2006.10.037

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

Zhou, G. X., Ge, Z., Dorwart, J., Izzo, B., Kukura, J., Bicker, G., et al. (2003). Determination and differentiation of surface and bound water in drying substances by near infrared spectroscopy. *J. Pharm. Sci.* 92, 1058–1065. doi: 10.1002/jps.10375

**Conflict of Interest Statement:** OH is an AbbVie employee and may own AbbVie stock/options.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 25 October 2018
doi: 10.3389/fchem.2018.00515

# Eliminating Non-linear Raman Shift Displacement Between Spectrometers via Moving Window Fast Fourier Transform Cross-Correlation

*Hui Chen*[1,2,3†], *Yan Liu*[1†], *Feng Lu*[1*], *Yongbing Cao*[2,4*] *and Zhi-Min Zhang*[5*]

[1] School of Pharmacy, Second Military Medical University, Shanghai, China, [2] Department of Vascular Disease, Shanghai TCM-Integrated Hospital, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3] Quality Control Department, Shanghai Diracarta Biomedical Technology Co., Ltd, Shanghai, China, [4] Department of Foundation and New Drug Research, Shanghai TCM-Integrated Institute of Vascular Disease, Shanghai, China, [5] College of Chemistry and Chemical Engineering, Central South University, Changsha, China

OPEN ACCESS

**Edited by:**
Hoang Vu Dang,
Hanoi University of Pharmacy, Vietnam

**Reviewed by:**
Andreas Borgschulte,
Swiss Federal Laboratories for Materials Science and Technology, Switzerland
Sebastian Primpke,
Alfred Wegener Institut Helmholtz Zentrum für Polar und Meeresforschung, Germany

***Correspondence:**
Feng Lu
fenglufeng@hotmail.com
Yongbing Cao
ybcao@vip.sina.com
Zhi-Min Zhang
zhangzhimin@csu.edu.cn

[†] These authors have contributed equally to this work and are co-first authors

**Specialty section:**
This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry

**Received:** 08 February 2018
**Accepted:** 05 October 2018
**Published:** 25 October 2018

**Citation:**
Chen H, Liu Y, Lu F, Cao Y and Zhang Z-M (2018) Eliminating Non-linear Raman Shift Displacement Between Spectrometers via Moving Window Fast Fourier Transform Cross-Correlation. Front. Chem. 6:515. doi: 10.3389/fchem.2018.00515

Obtaining consistent spectra by using different spectrometers is of critical importance to the fields that rely heavily on Raman spectroscopy. The quality of both qualitative and quantitative analysis depends on the stability of specific Raman peak shifts across instruments. Non-linear drifts in the Raman shifts can, however, introduce additional complexity in model building, potentially even rendering a model impractical. Fortunately, various types of shift correction methods can be applied in data preprocessing in order to address this problem. In this work, a moving window fast Fourier transform cross-correlation is developed to correct non-linear shifts for synchronization of spectra obtained from different Raman instruments. The performance of this method is demonstrated by using a series of Raman spectra of pharmaceuticals as well as comparing with data obtained by using an existing standard Raman shift scattering procedure. The results show that after the removal of shift displacements, the spectral consistency improves significantly, i.e., the spectral correlation coefficient of the two Raman instruments increased from 0.87 to 0.95. The developed standardization method has, to a certain extent, reduced instrumental systematic errors caused by measurement, while enhancing spectral compatibility and consistency through a simple and flexible moving window procedure.

**Keywords: Raman instruments, shift correction, cross-correlation, fast fourier transform, moving window**

## INTRODUCTION

Over the last few decades, the use of Raman spectroscopy in combination with chemometric methods has increased significantly for analysis of pharmaceutical products (Sacré et al., 2010; Dégardin et al., 2011; Loethen et al., 2015), detection of food adulteration (Zou et al., 2009; Cheng et al., 2010), and other applications (Mrozek et al., 2004; Taleb et al., 2006; Muehlethaler et al., 2011). Raman spectroscopy is a powerful tool for sample analysis and benefits from several advantages such as high speed, simplicity, non-destructive nature, and cost-effectiveness. To date,

it has been extensively applied in pharmaceutical analysis by constructing multivariate calibration models. However, these models will be invalid if an existing calibration model is applied to spectra that are collected on a different occasion or a separate instrument, or when the response of an old instrument suffers from variations (Du et al., 2011; Brown, 2013). These variations may, if left untreated, dominate the calibration models, thereby making analysis of samples impractical. Consequently, chemometric techniques have been used to circumvent these problems through instrumental transfer or standardization so as to isolate and compensate for any instrumental and environmental variations.

A number of methods, including both instrumental transfer and standardization, have been discussed in the literature (Wang et al., 1991, 1992; De Noord, 1994; Mann and Vickers, 1999; Nguyen Quang et al., 1999; Hutsebaut et al., 2005; Kompany-Zareh and van den Berg, 2010; Rodriguez et al., 2011b; Weatherall et al., 2013). The direct standardization (DS) and piecewise direct standardization (PDS) developed by Wang et al. (1991, 1992) are the most extensively used procedures for spectral response standardization. Using the PDS method, Gryniewicz-Ruzicka (Gryniewicz-Ruzicka et al., 2011) obtained a very low detection limit for diethylene glycol in pharmaceutical-grade glycerin by using five portable Raman spectrometers. This method, however, requires the user to measure several standards prior to analyzing samples. In addition, both the use of the moving window strategy and the selection of principal components have a noticeable impact on the performance of PDS, which needs to be determined carefully. Furthermore, neither the DS nor the PDS method can deal with different (i.e., non-linear) shifts in the peaks in Raman spectra. It is worth mentioning that in contrast to the various instrumental spectral responses, Raman shift inconsistencies arise mainly from different charge-coupled device (CCD) detectors (Vickers and Mann, 1999). Nonetheless, the use of inconsistent spectra will diminish significantly the predictive power of a calibration model. As a result, the removal of Raman shifts or wavelength inconsistencies for spectra synchronization has become a particularly significant aspect of Raman spectroscopy analysis. In 1996, a mathematical procedure to correct wavelength drifts to synchronize Raman spectra was presented by Booksh et al. (1996). Typically, empirical data are required to select a number of principal components and channels to increase the synchronization precision. Westad and Martens (1999) developed a more general concept of shift determination and tested it on Raman spectra. The results revealed, however, that the spectra were not reproduced exactly after removal of peak drifts exceeding a discrete spectral resolution. Hutsebaut et al. (2005) used a Raman shift standard scattering (SSS for short) method in combination with a linear fitting to determine shift drifts between measured Raman peak and reference positions. A similar approach was used by Rodriguez et al. (2011b) to transfer Raman spectral libraries among instruments. Nevertheless, the use of Raman shift standards is inappropriate for in-line monitoring applications as a result of the difficulties associated with incorporating one or more of the materials proposed as shift standards in a system for in-line measurements. Recently, another approach for the removal of disturbing factors in the CCD responses and instrumental apparatus functions was proposed by Weatherall et al. (2013). Unfortunately, the use of *baselineWavelet* continuous wavelet transform as a function to identify major peaks' positions accurately requires idealized line profiles of the corresponding peaks, which is not practical for real Raman spectra. In addition, several parameters that influence the final results, such as the width of the window and the choice of the signal-to-noise threshold, need to be specified, mostly by the users.

As a result of the multifarious theoretical and practical limitations of the existing instrument standardization methods (Chen et al., 2015), there is a significant demand for methods that are easier to implement (i.e., fewer or even no tunable parameters required) in order to acquire better analytical performance. Accordingly, we introduced a cross-correlation method in order to address the problems (such as tunable parameters, need idealized line profiles, etc.) discussed above. Generally, in signal processing, cross-correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them, and is also known as a sliding dot product or sliding inner-product (Welch, 1974; Goshtasby et al., 1984). When coupled with fast Fourier transform (FFT) algorithms, the efficiency of FFT can be exploited in the numerical computation of cross-correlations, accelerating thus the convolution calculation (Bracewell, 1980). FFT cross-correlation may therefore be the fastest method in signal processing for shift correction (Bergland, 1969), and benefits from many advantages such as high speed and accuracy. Moreover, it also eliminates the requirement for alignment parameters. Previously, two alignment methods were proposed to estimate the shifts between segments in large chromatographic and spectral datasets, namely, peak alignment by FFT (Wong et al., 2005b) and recursive alignment by FFT (Wong et al., 2005a). However, these two methods move segments by insertion and deletion of data points at the start and end of segments, without considering peak information, which may cause changes in the shapes of peaks by introducing artifacts and removing peak points. Zhan et al. (Zhang et al., 2012) developed another method, known as the multi-scale peak alignment (MSPA) method, to synchronize peaks against a reference chromatogram (aligning peaks from large to small scales), which is accelerated by the application of FFT cross-correlation while preserving peak shape during synchronization. Similarly, Li et al. (2013a) developed a moving window FFT cross-correlation (MWFFT) method to effectively synchronize high-throughput chromatograms without segment size optimization. However, the Raman spectra profiles were different from the chromatograms, which required peak fitting to obtain perfect profiles and a precise Raman shift.

In the present work, the MWFFT was improved and subsequently applied to spectral standardization to address the issues associated with spectral drifts in Raman spectrometers. The performance of this method was compared to that of the SSS method (Hutsebaut et al., 2005) by using two Raman datasets from primary and secondary spectrometers. The aim of our study was to make the MWFFT as a powerful and practical method for standardization across

Raman spectrometers, which can be easily implemented and well-suited for solving Raman shifts displacements between spectrometers.

# MATERIALS AND METHODS

## Standards and Samples

Standards (acetaminophen and cyclohexane) were provided by the National Institute for the Control of Pharmaceutical and Biological Products. Pharmaceutical tablets (listed in **Table 1**) from five different manufacturers were provided by the Shanghai Institute for Food and Drug Control.

## Raman Spectrometers

Two Raman instruments with an excitation wavelength of 785 nm were used, and their physical parameters are listed in **Table 2**. In this work, the i-Raman is regarded as the "master" (primary) instrument, while the GemRam is regarded as the "slave" (secondary) instrument.

The integration times of the standards and drugs were of 2 and 3 s, respectively. Unless stated otherwise, six Raman spectra were collected for each drug during the experiment. It is worth noting that the final spectrum of each drug was calculated as the average of spectra collected from a variety of positions. Moreover, only the spectral region containing the most abundant information (i.e., 300–1,700 $cm^{-1}$) was used in subsequent data analysis.

## Cross-Correlation

In signal processing, cross-correlation is a standard technique to calculate the similarity between and estimate the linear shift of two signals as a function of one relative to the other, which is also known as the sliding dot product. It is obvious that any changes involving the shifting of one signal will affect the correlation coefficient calculated for any combination of two signals that includes this shifted signal. For two discrete signals such as those in the Raman spectra, the cross-correlation is defined as:

$$\mathbf{c}(j) = \sum_i \left(\mathbf{r}(i) - \bar{r}\right)\left(\mathbf{s}(i+j) - \bar{s}\right) \Big/ \sqrt{\sum_i \left(\mathbf{r}(i) - \bar{r}\right)^2} \sqrt{\sum_i \left(\mathbf{s}(i+j) - \bar{s}\right)^2} \quad (1)$$

where r is the reference signal, s is the signal to be synchronized, c is the cross-correlation values for all lags. As a simple example, consider two simulated Raman spectra *r* and *s* that differ only by a known displacement of 90 points along the x-axis. We can determine by how much *s* be shifted along the x-axis in order to maximize its similarity to *r* by using cross-correlation. The above formula slides *s* along the x-axis, calculating the sum of their product at each position. When the value of *c* is maximized, i.e., the signals match well due to peak synchronization, they make the most significant contribution to the sum of their product. A visual description of the calculation procedure of cross-correlation and estimation of shifts between signals via cross-correlation is shown in **Figure 1**.

**TABLE 1 |** Correlation coefficients of drug tablets before and after shift correction.

| Drugs | Batches | i-Raman & GemRam | | |
|---|---|---|---|---|
| | | $r^u$ | $R^s$ | $R^m$ |
| Acyclovir tablets | 20100301 | 0.9424 | 0.9893 | 0.9905 |
| | 20120102 | 0.9430 | 0.9888 | 0.9906 |
| | 130302 | 0.9186 | 0.9567 | 0.9570 |
| | 20111201 | 0.9172 | 0.9593 | 0.9605 |
| | 20101102 | 0.9027 | 0.9585 | 0.9619 |
| | 20110501 | 0.9358 | 0.9903 | 0.9932 |
| | 20101103 | 0.9356 | 0.9906 | 0.9938 |
| | 20100901 | 0.9422 | 0.9923 | 0.9944 |
| | 20110401 | 0.9377 | 0.9914 | 0.9942 |
| | 20120101 | 0.9398 | 0.9914 | 0.9937 |
| | 100301R | 0.9435 | 0.9924 | 0.9944 |
| | 090601P | 0.9365 | 0.9899 | 0.9927 |
| | 110101 | 0.9479 | 0.9932 | 0.9945 |
| | 100101P | 0.9477 | 0.9915 | 0.9923 |
| | 091101P | 0.9489 | 0.9900 | 0.9904 |
| Captopril tablets | 20101009 | 0.8964 | 0.9597 | 0.9644 |
| | 090406 | 0.9178 | 0.9672 | 0.9681 |
| | 63120501 | 0.8776 | 0.9642 | 0.9692 |
| | 110804 | 0.8919 | 0.9604 | 0.9644 |
| | 63120401 | 0.8769 | 0.9592 | 0.9620 |
| | 110202 | 0.9131 | 0.9666 | 0.9680 |
| | 63111001 | 0.9078 | 0.9770 | 0.9796 |
| | 110805 | 0.9196 | 0.9725 | 0.9747 |
| | 090404 | 0.9093 | 0.9658 | 0.9684 |
| | 121003 | 0.8918 | 0.9668 | 0.9720 |
| | 63120301 | 0.8843 | 0.9621 | 0.9642 |
| | 110901 | 0.9096 | 0.9489 | 0.9509 |
| | 090307 | 0.8923 | 0.9535 | 0.9558 |
| | 20101006 | 0.8970 | 0.9663 | 0.9714 |
| | 63110702 | 0.8704 | 0.9490 | 0.9522 |
| | 110801 | 0.9062 | 0.9789 | 0.9804 |
| | 20110515 | 0.8954 | 0.9669 | 0.9679 |
| | 20101005 | 0.9060 | 0.9517 | 0.9518 |
| | 110506 | 0.9213 | 0.9546 | 0.9560 |
| | 110702 | 0.9311 | 0.9701 | 0.9717 |
| | 110903 | 0.9271 | 0.9573 | 0.9582 |
| | 110804 | 0.9181 | 0.9543 | 0.9564 |
| | 110604 | 0.9274 | 0.9649 | 0.9651 |
| | 110803 | 0.9084 | 0.9506 | 0.9512 |
| | 20101004 | 0.9254 | 0.9503 | 0.9503 |

*[u]Correlation coefficient before shift correction; [s]Correlation coefficient by SSS; [m]Correlation coefficient by MWFFT.*

## Moving Window FFT Cross-Correlation

The FFT is typically used to calculate the cross-correlation between 1D and 2D signals (Papoulis, 1962; Cooley et al., 1969; Dutt and Rokhlin, 1993). In the present work, FFT was used to increase the speed of cross-correlation between two datasets, in which one signal may be shifted relative

**TABLE 2 |** Physical parameters for the two Raman spectrometers used in this work.

| Spectrometer | Manufacturer | Laser power (mw) | Spectral range ($cm^{-1}$) | System resolution ($cm^{-1}$) | CCD pixel number |
|---|---|---|---|---|---|
| i-Raman | B&W Tek Inc | 100 | 175–2700 | 3 | 2048 |
| GemRam | B&W Tek Inc | 100 | 175–2700 | 3.5 | 2048 |

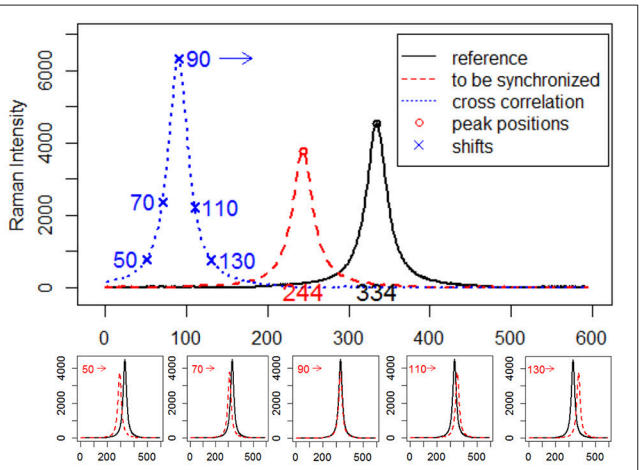



**FIGURE 1 |** Estimation of displacements between simulated Raman spectra by cross-correlation.

to another. In addition, and perhaps more significantly for its application to the spectral synchronization problem, FFT cross-correlation is not heuristic and thus can identify consistently the best match between signals by finding the maximum correlation coefficient (Wong et al., 2005b).

Usually, the cross-correlation method can only estimate linear shifts between Raman spectra. However, Raman shift displacements are often non-linear in real samples. Consequently, we adopted the moving window procedure in this work to address this problem. In this procedure, the shifts relative to the reference can be estimated by FFT cross-correlation, allowing us to obtain the shift profiles of all samples. Furthermore, MWFFT can be implemented and optimized simply and effectively only if a moving window of appropriate size is utilized. With a window moving from the beginning to the end of the two spectra, one can obtain a matrix of shift points. Accordingly, the shift profile can be obtained by calculating the mode value of each column of the shift matrix. **Figure 2A** shows an example Raman shift profile estimated by using the moving window strategy and FFT cross-correlation. It is apparent from the obtained shift profile that non-linear shifts exist across the entire spectral region, while the change points are observed in two regions with different shifts. By moving the continuous region around the change points, the synchronization procedure can be finished smoothly to obtain the synchronized spectrum, which can be seen in **Figure 2B**, with all the non-linear shifts successfully synchronized.

# RESULTS

There are two common ways of correcting the x-axis in Raman spectrometers (McCreery, 2005). The first one is to simply use the SSS method (Hutsebaut et al., 2005); the second one is based on absolute frequency calibration using the emission line spectra of gases. The SSS method, which requires the acquisition of Raman spectra of common materials with well-established Raman shift peak frequencies in order to correct the Raman shift axis directly, is used as a comparative method in this work. Several well-known Raman shift chemical standards, namely, cyclohexane and acetaminophen, are chosen over others for this study since their spectral combination can provide more signals in the region from 300 to 1,700 $cm^{-1}$ (see **Table 3**). The left panel in **Figure 3** shows the spectra acquired for the used chemical standards on two instruments, while the right panel shows a plot of their differences. It should be noted that when the SSS method was used, the spectra acquired on the primary instrument were regarded as the reference, i.e., the peak positions in these spectra were used for synchronization. The relevant peak positions obtained on the secondary instrument are compared to those obtained on the primary instrument and are subsequently subtracted from the primary peak positions to afford the corresponding shift displacements. Linear fitting is then used to describe the shift displacements between the two instruments. Finally, the shift correction is carried out by linear interpolation.

## Synchronization of Pharmaceutical Datasets

Data synchronization of the raw Raman spectra are presented to evaluate the performance of the MWFFT method (**Figure 2**). In order to gain further insight into the two shift correction algorithms, and the properties and advantages of MWFFT in particular, different batches of pharmaceutical tablets were examined to verify the practicability and effectiveness of MWFFT. **Figure 4** describes the application of MWFFT—each tablet from a total of 40 drugs was analyzed on average six times on two instruments to obtain six different spectra. Subsequently, these spectra were detected for outlier. The average spectrum obtained from six spectra acquired on the primary instrument can be regarded as a reference without outliers. Analogously, we obtained the spectrum of the same tablet on the secondary instrument, and this represents the spectrum to be synchronized. Finally, MWFFT was applied to remove the shift displacements in order to synchronize the spectra across the two instruments.

Prior to data analysis, adaptive iteratively reweighted penalized least squares (airPLS) (Zhang et al., 2010a,b; Li et al.,

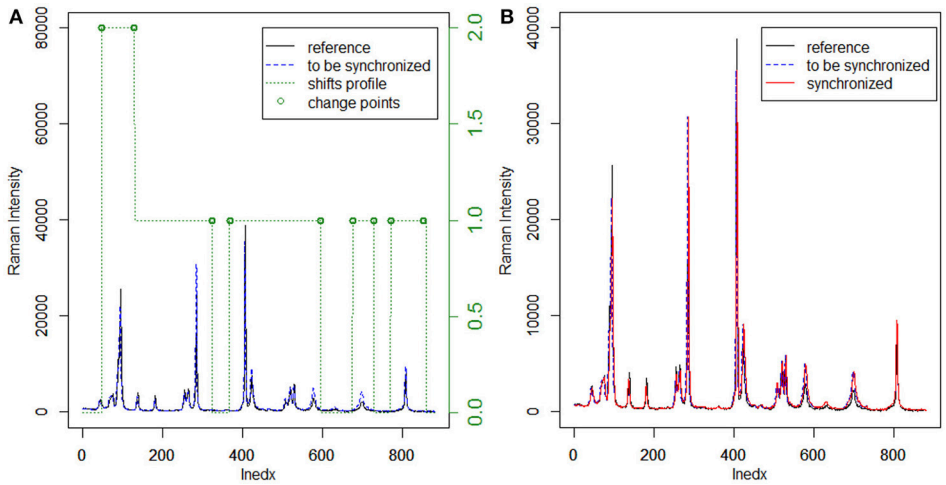



**FIGURE 2 |** Application of MWFFT to synchronize Raman spectra: **(A)** estimation of nonlinear shift profile between two spectra. **(B)** The continuous regions are moved around change points to obtain the synchronized spectrum.

2013b) baseline correction and Savitzky–Golay smoothing (Savitzky and Golay, 1964) (a 9-point wide window and a second-order polynomial) were used in the preprocessing of a variety of pharmaceutical datasets. All processing tasks were implemented on a personal computer (CPU: 2.53G, RAM: 8GB) with MATLAB R2013a. Firstly, we demonstrate the effect of MWFFT by using the pharmaceutical datasets (**Figure 5**). The primary instrument spectra (black lines) are used as references for synchronization. **Figure 5** shows the magnified versions of the sample profiles, focusing on a particular set of peaks in order to allow the performance of the MWFFT method to be evaluated by visual inspection. For the acyclovir and captopril datasets, it can be seen that before synchronization (top panel in **Figure 5**), the peaks in the spectrum collected on the secondary instrument are de-synchronized with respect to that obtained on the primary instrument, and vary from sample to sample. After synchronization (middle panel in **Figure 5**) using the MWFFT method, it is apparent that all the spectra are now properly synchronized. This outcome is attributed to the action of the MWFFT method, which appropriately slides the peaks to match the reference spectrum with a window size of 70 points. In addition, for the sake of comparison, the results obtained using the SSS method for the same spectra are displayed in the bottom panel of **Figure 5**.

## Correlation Coefficient After Synchronization

The correlation, or distance, between a signal and the reference point is often used as an optimization objective function—when the signals match, the correlation coefficient is maximized. In this case, correlation coefficient can be used to assess the synchronization problem (Lee Rodgers and Nicewander, 1988). Generally, the correlation coefficient is a good descriptor of similarity, with a value of 1.00 indicating a perfect match, while 0 indicates significant dissimilarity. The correlation coefficient is simple to use and possesses several desirable properties, which we discussed in detail in our previous work (Gao et al., 2014). The correlation coefficient can be calculated by using the following equations:

$$r = \sum_{i=1}^{n}(X_i^p - \bar{x}^p)(X_i^s - \bar{x}^s) \Big/ \sqrt{\sum_{i=1}^{n}(X_i^p - \bar{x}^p)^2 \sum_{i=1}^{n}(X_i^s - \bar{x}^s)^2} \quad (2)$$

$$R = \sum_{i=1}^{n}(X_i^p - \bar{x}^p)(X_i^{sa} - \bar{x}^{sa}) \Big/ \sqrt{\sum_{i=1}^{n}(X_i^p - \bar{x}^p)^2 \sum_{i=1}^{n}(X_i^{sa} - \bar{x}^{sa})^2} \quad (3)$$

Here, $X^p$ and $X^s$ represent the spectra of n drugs measured on the primary and secondary instruments, respectively. Parameters $x^p$, $\bar{x}^s$, and $\bar{x}^{sa}$ represent the average spectra of $X^p$, $X^s$, and $X^{sa}$, respectively. $X^{sa}$ indicates the secondary shift corrected spectrum, while r and R denote a similarity between the primary original spectrum and the secondary spectrum (before or after shift correction). The correlation coefficient of each drug's spectrum was calculated, and the results are summarized in **Table 1**. During the preprocessing, linear interpolation was used to re-compute intensity based on the master Raman shift x-axis in order to unify the spectra obtained using the primary and secondary instruments. It is apparent from **Table 1** that the correlation coefficients between the two instruments improved significantly after shift correction.

As can be seen in **Table 1**, the correlation coefficient assessment prior to the shift correction exhibited a slight variation among different batches of a drug. Nevertheless, these variations are within the three-sigma range. Portable spectrometers are often based on the use of library-based

spectral correlation methods (Carron and Cox, 2010), which frequently utilize the hit-quality index (HQI) as the figure of merit to characterize the correlation with each other. The typical minimum threshold that classifies an unknown sample as a "Pass" is 0.95 (Rodriguez et al., 2011a, 2013), which is similar to the correlation coefficient. Clearly, the MWFFT method makes a significant contribution to the level of similarity for the spectra obtained using the slave instrument. The synchronization increased the similarities for all drugs above the verification threshold of 0.95, while the similarity for one captopril tablet remained under 0.95 when the SSS procedure was used. Consequently, it is obvious that the MWFFT method can correct the non-linear shifts successfully, synchronizing thus the secondary spectra to the reference spectra in a time-effective manner. In addition, MWFFT can reduce the systematic differences across spectrometers, which can increase the spectral consistency of different instruments as well as the compatibility with library search. Furthermore, this method can be used as an on-line standardization method across Raman instruments in the future.

**TABLE 3 |** Raman shifts ($cm^{-1}$) used to calibrate standard samples.

| Standard | Raman shift (± standard deviation)[a] | |
|---|---|---|
| **4-ACETAMIDOPHENOL** | | |
| | 329.2 ± 0.5 | 1168.5 ± 0.6 |
| | 390.9 ± 0.8 | 1236.8 ± 0.5 |
| | 465.1 ± 0.3 | 1278.5 ± 0.5 |
| | 504.0 ± 0.6 | 1323.9 ± 0.5 |
| | 651.6 ± 0.5 | 1371.5 ± 0.1 |
| | 710.8 ± 0.7 | 1515.1 ± 0.7 |
| | 797.2 ± 0.5 | 1561.5 ± 0.5 |
| | 834.5 ± 0.5 | 1648.4 ± 0.5 |
| | 857.9 ± 0.5 | 1278.5 ± 0.5 |
| | 968.7 ± 0.6 | 1168.5 ± 0.6 |
| | 1105.5 ± 0.3 | 1236.8 ± 0.5 |
| **CYCLOHEXANE** | | |
| | 384.1 ± 0.8 | 1157.6 ± 0.9 |
| | 426.3 ± 0.4 | 1266.4 ± 0.6 |
| | 801.3 ± 0.96 | 1444.4 ± 0.3 |
| | 1028.3 ± 0.5 | 384.1 ± 0.8 |

[a] *Values as reported by ASTM E1840-96.*

## DISCUSSION

### Selection of Reference Spectrum

A wide application of the MWFFT method necessitates the selection of an appropriate reference spectrum. When a drug sample is measured on a secondary instrument to obtain an average spectrum for synchronization, its corresponding standard spectrum contained in the existing spectral library can

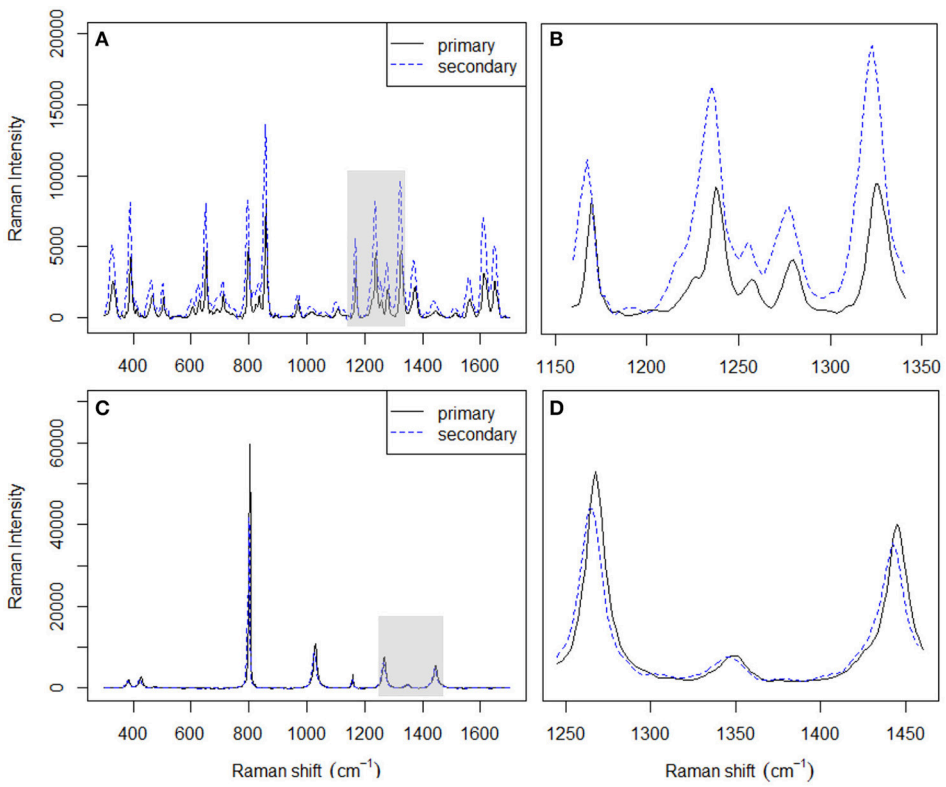



**FIGURE 3 |** Spectra of acetaminophen **(A)** and cyclohexane **(C)** acquired on two different instruments. Magnified spectral differences in **(B,D)** correspond to the shaded areas in **(A,C)**, respectively.

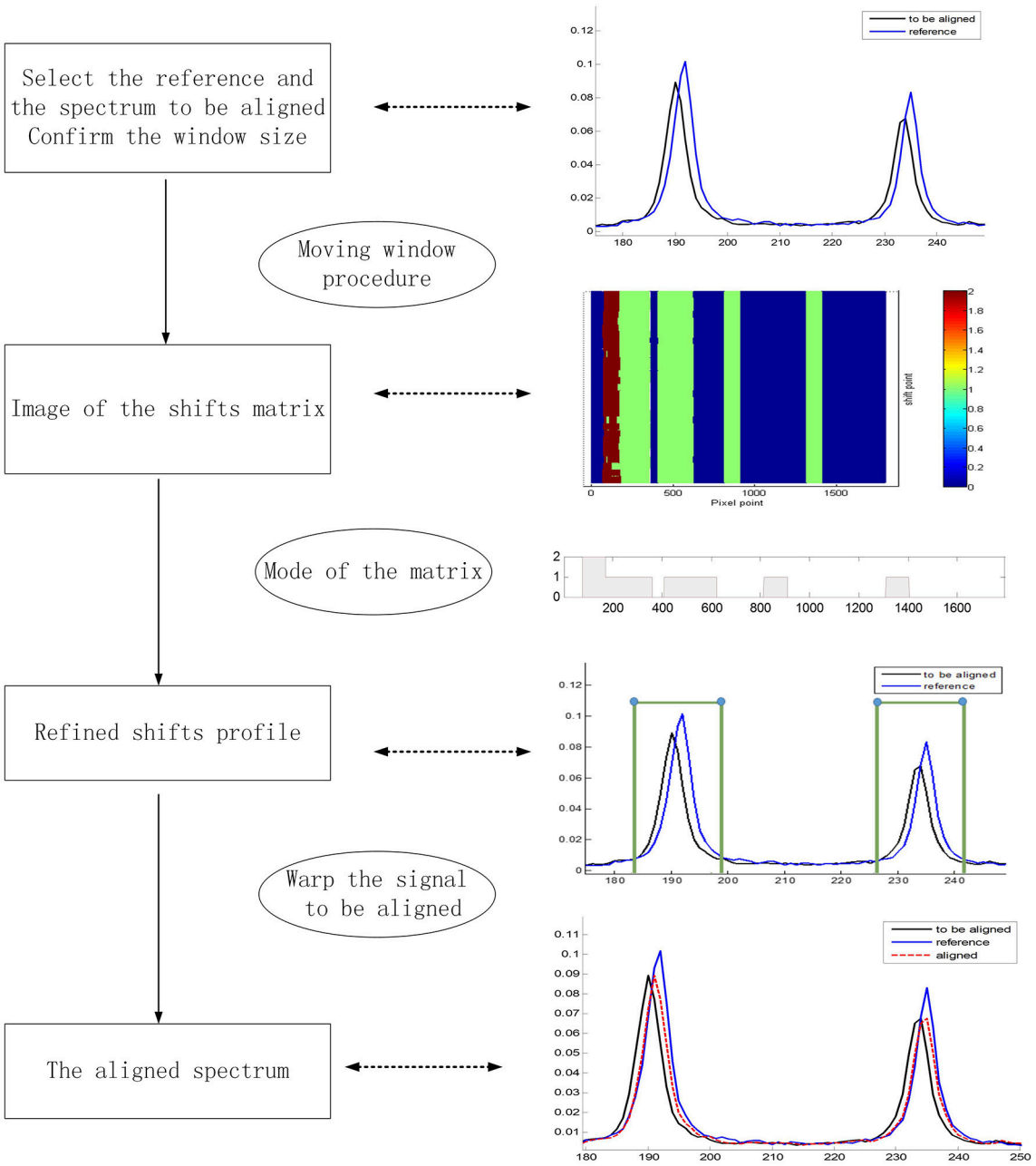



**FIGURE 4 |** Flowchart describing the framework of MWFFT.

be certainly used as the reference to correct shift displacements. However, when the spectral library does not contain the required reference spectrum, it would be preferable to use the reference spectra of existing drugs with the same generic name in the database in order to obtain a new matrix of shift points. As a result, the shift profile of the new drug can be calculated from the mode of each column of the matrix. Through this profile, one can obtain a new reference spectrum by shift correction, which can be subsequently applied. Otherwise, one can regard the new sample spectrum directly as a reference, and save it in the database for subsequent analysis. The entire procedure is depicted in **Figure 6**.

## Avoiding Peak Detection Using the Moving Window Strategy

The existing peak detection methods, e.g., the wavelet and ridge line peak picking method, need idealized line profiles of the corresponding peaks in order to detect the displacements accurately, which is not practical for the spectroscopic analysis of real samples. Moreover, several parameters need to be specified

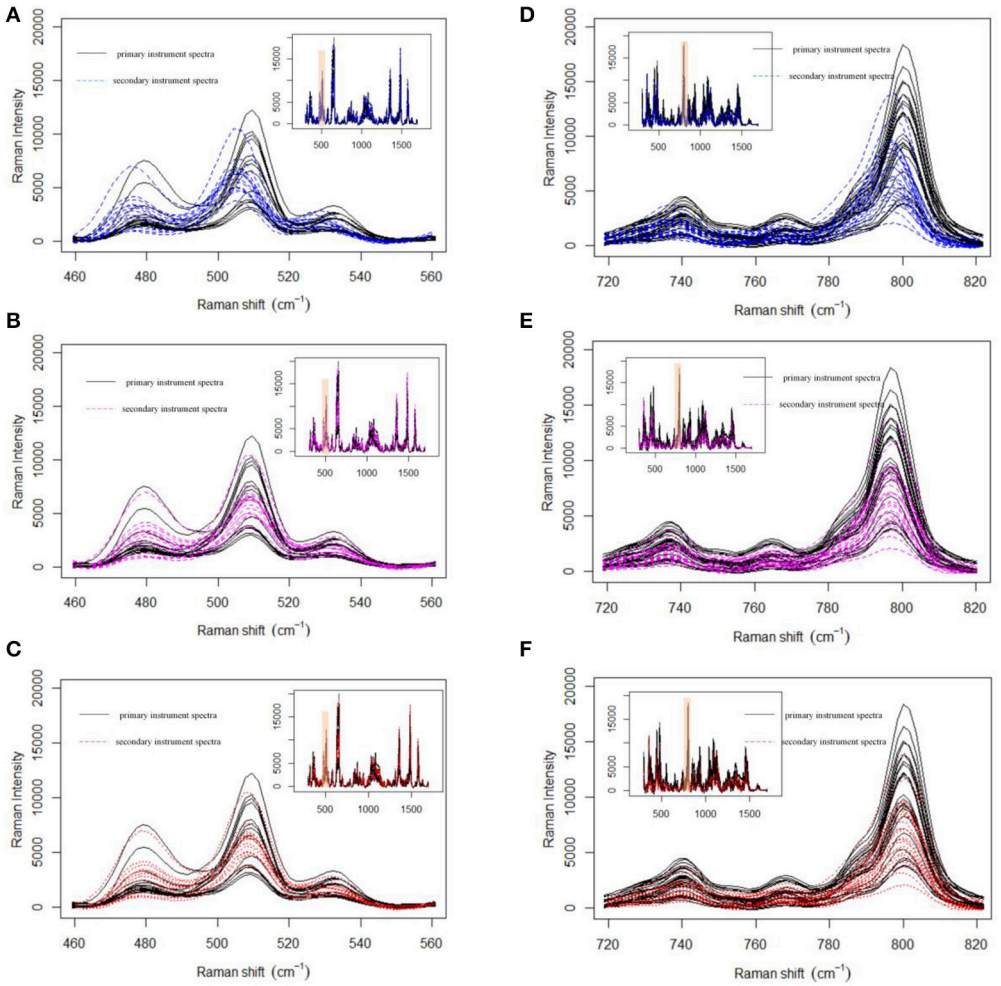



**FIGURE 5 |** Shift correction data for both acyclovir and captopril datasets with MWFFT and SSS: **(A–C)** acyclovir dataset **(A)** before synchronization, **(B)** synchronized by MWFFT, and **(C)** synchronized by SSS; **(D–F)** captopril dataset **(D)** before synchronization, **(E)** synchronized by MWFFT, and **(F)** synchronized by SSS. The black lines indicate the reference spectra. The inset shows the full Raman spectra, whereas the shaded areas indicate the region magnified in the main panel.

with a priori knowledge, which largely influence the final results and can be difficult to implement in C programming language. By contrast, the use of the moving window strategy can allow an estimation of non-linear shifts between spectra flexibly and without peak detection for peak synchronization. With a window that moves from the beginning to the end of two spectra, one can obtain an N-dimensional matrix of shift points, where the data points of a Raman spectrum are N. In this case, the shift profile can be calculated from the mode of each column of the matrix, while the mean and median of the matrix can outline the paths of the shifts. The Raman shift profile of metronidazole tablet is shown in **Figure 2A** using a green dotted line. It is apparent that the profiles in all regions are corrected by the MWFFT method, meaning this method is sufficiently flexible for estimation of non-linear shifts between spectra.

## Advantages of the MWFFT Method

The MWFFT method has several distinctive advantages when compared to the traditional methods as a result of the continuity and redundancy of the moving window procedure. Usually, the direct evaluation of cross-correlation requires O ($N^2$) time complexity for a Raman spectrum of length N, which is time-consuming for spectra with thousands of data points. Fortunately, cross-correlation can be calculated by using FFT much more efficiently since it can significantly decrease the time complexity of cross-correlation from O ($N^2$) to O (NlogN). The use of the moving window strategy with FFT cross-correlation, with a window size $w$, leads to a time complexity of one window $w\log w$. Accordingly, the time complexity of MWFFT is $Nw\log w$, where N represents the number of data points in a Raman spectrum.

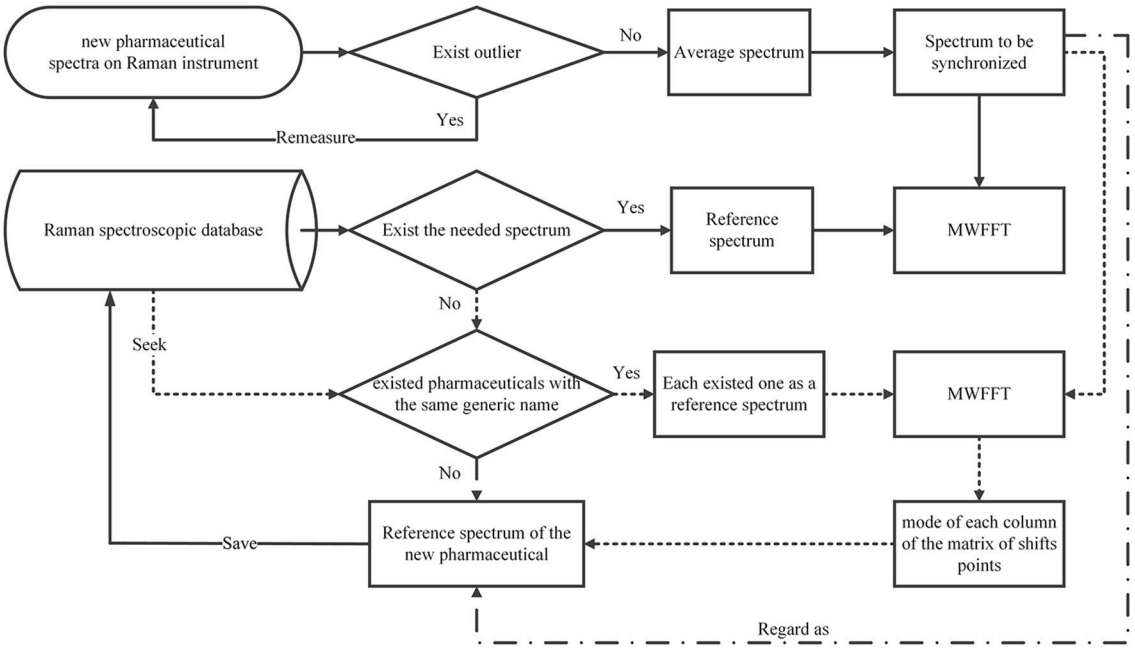



**FIGURE 6 |** Flowchart of selecting a new drug reference spectrum.

The MWFFT method evaluates the shift of each point. In the moving window strategy, only one parameter needs to be taken into account, which makes this procedure simple and practical, as there is no need for chemical standards. By contrast, the SSS method requires the use of some chemical standards in order to locate the position of each peak, which is used in turn to obtain the corresponding shift displacement. After the shift of each point is estimated by MWFFT, the points in the spectrum are shifted according to their shifts by insertion and deletion. The present work introduced a change point, i.e., a discontinuity point in the shift profile. It is possible to see that the change points (**Figure 2A**), around which insertions and deletions occur frequently, are not in the peak region. Consequently, peak distortions can be effectively avoided, allowing the peak shape to be preserved during the synchronization procedure with MWFFT. Overall, the advantages associated with the use of non-linear shift estimation, insertion and deletion around change points, and shape preservation make MWFFT a flexible, rapid, practical, and precise method for correcting shifts in synchronization of Raman datasets.

## Evaluation of the Synchronization Quality

Generally, Raman spectra will become more consistent, exhibit higher correlation coefficients, and be more similar to each other after a successful synchronization. The correlation coefficient can be used as a criterion for assessing the synchronization quality between the primary and secondary spectra. The synchronized spectra are commonly used to perform library-based searches and are further analyzed by chemometric algorithms. Usually, distance and Euclidean distance in particular (Juday, 1993), can also be a good criterion for evaluating the quality of synchronization. Generally, the more similar the spectra are, the smaller is the Euclidean distance between them, and vice versa. In this work, the mean Euclidean distance ($D_{mean}$) is calculated as follows:

$$D_{mean} = {}^{1}/_{n} \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{k} (X_{i,j}^{p} - X_{i,j}^{s})^{2}} \quad (4)$$

where the rows of matrix X correspond to observations ($n$), while the columns correspond to variables ($k$). $X_i^p$ and $X_i^s$ are the $i$th primary (reference) spectrum and secondary spectrum, respectively. It is worth mentioning at this stage that the normalization algorithm (Heraud et al., 2006) is used to scale the spectra within a similar range before calculating the distances. The results are summarized in **Table 4**. It is apparent that the mean Euclidean distance of the

**TABLE 4 |** Mean Euclidean distances of the used drug datasets shift corrected by SSS and MWFFT.

| Datasets | Shift correction methods | | |
|---|---|---|---|
| | Uncorrected | SSS | MWFFT |
| $D_{mean}{}^{ac}$ | 1.9222 | 1.1637 | 1.1488 |
| $D_{mean}{}^{ca}$ | 2.5565 | 1.8900 | 1.8798 |

*[ac]Mean Euclidean distances of acyclovir datasets; [ca]Mean Euclidean distances of captopril datasets.*

pharmaceutical datasets shift-corrected by SSS and MWFFT were considerably reduced when compared to the uncorrected ones. In addition, for the two datasets, MWFFT performed slightly better than the SSS method in terms of non-linear shift correction.

## CONCLUSIONS

Methods for the synchronization of spectra are indispensable for successful applications using different spectrometers. In the present work, we used the moving window strategy in combination with FFT cross-correlation to synchronize Raman spectra. This technique, abbreviated as MWFFT, was shown to eliminate accurately and effectively non-linear shift displacements between Raman spectra. Owing to the continuity of the moving window technique, non-linear shifts are corrected and shift profiles are obtained for each spectrum. In general, the use of the FFT cross-correlation methodology is time-saving and results in a significant improvement in speed. Moreover, this method can reduce or even remove systematic differences between Raman spectrometers (a dramatic increase in similarity from 0.87 to 0.95 after synchronization of the spectra between master (primary) and slave (secondary) spectrometers), as well as the compatibility with Raman spectral library. It is better than the SSS method in terms of correcting non-linear shifts and does not require the use of Raman shift standards. These advantages make MWFFT a promising shift correction method that addresses the demand for automated, flexible, rapid, and reliable data preprocessing, which plays an important role in Raman spectroscopy analysis using different spectrometers. Finally, MWFFT can be easily implemented with C and C++ programming languages (available as open source package at http://code.google.com/p/mwfft), which may be well-suited to solving the Raman shift displacements between spectrometers in the fields that rely heavily on the use of Raman spectrometers.

## ETHICS STATEMENT

The experimental protocol was approved by the Research Ethics Committee of The Second Medical University and Shanghai University of Traditional Chinese Medicine, The findings and conclusions in this article have not been formally disseminated by the State Food and Drug Administration and should not be construed to represent any agency determination or policy.

## AUTHOR CONTRIBUTIONS

HC designed and carried out experiments. Z-MZ and YL assisted with analyzing the results and discussions. HC and Z-MZ wrote the manuscript. FL reviewed and edited the manuscript. YC reviewed and checked our manuscript, gave constructive amendments to the text, and also approved the version to be published.

## FUNDING

This work is financially supported by Ministry of Science and Technology of the People's Republic of China (2017YFF0210103, 2012YQ180132). The studies meet with the approval of the university's review board.

## REFERENCES

Bergland, G. (1969). A guided tour of the fast Fourier transform. *Spectr. IEEE* 6, 41–52. doi: 10.1109/MSPEC.1969.5213896

Booksh, K. S., Stellman, C. M., Bell, W. C., and Myrick, M. L. (1996). Mathematical alignment of wavelength-shifted optical spectra for qualitative and quantitative analysis. *Appl. Spectrosc.* 50, 139–147. doi: 10.1366/0003702963906500

Bracewell, R. N. (1980). *Fourier Transform and its Applications*. New York, NY: European Journal of Operational Research.

Brown, S. D. (2013). "Transfer of multivariate calibration models," in *Reference Module in Chemistry, Molecular Sciences and Chemical Engineering* (Waltham, MA: Elsevier), 345–378.

Carron, K., and Cox, R. (2010). Qualitative analysis and the answer box: a perspective on portable Raman spectroscopy. *Analyt. Chem.* 82:3419. doi: 10.1021/ac901951b

Chen, H., Zhang, Z. M., Miao, L., Zhan, D. J., Zheng, Y. B., Liu, Y., et al. (2015). Automatic standardization method for Raman spectrometers with applications to pharmaceuticals. *J. Raman Spectrosc.* 46, 147–154. doi: 10.1002/jrs.4602

Cheng, Y., Dong, Y., Wu, J., Yang, X., Bai, H., Zheng, H., et al. (2010). Screening melamine adulterant in milk powder with laser Raman spectrometry. *J. Food Composit. Anal.* 23, 199–202. doi: 10.1016/j.jfca.2009.08.006

Cooley, J. W., Lewis, P. A., and Welch, P. D. (1969). The fast Fourier transform and its applications. *Educ. IEEE Trans.* 12, 27–34. doi: 10.1109/TE.1969.4320436

De Noord, O. E. (1994). Multivariate calibration standardization. *Chemometr. Intell. Lab. Syst.* 25, 85–97. doi: 10.1016/0169-7439(94)85037-2

Dégardin, K., Roggo, Y., Been, F., and Margot, P. (2011). Detection and chemical profiling of medicine counterfeits by Raman spectroscopy and chemometrics. *Anal. Chim. Acta* 705, 334–341. doi: 10.1016/j.aca.2011.07.043

Du, W., Chen, Z. P., Zhong, L. J., Wang, S. X., Yu, R. Q., Nordon, A., et al. (2011). Maintaining the predictive abilities of multivariate calibration models by spectral space transformation. *Anal. Chim. Acta* 690, 64–70. doi: 10.1016/j.aca.2011.02.014

Dutt, A., and Rokhlin, V. (1993). Fast Fourier transforms for nonequispaced data. *SIAM J. Sci. Comp.* 14, 1368–1393. doi: 10.1137/0914081

Gao, Q., Liu, Y., Li, H., Chen, H., Chai, Y., and Lu, F. (2014). Comparison of several chemometric methods of libraries and classifiers for the analysis of expired drugs based on Raman spectra. *J. Pharm. Biomed. Anal.* 94, 58–64. doi: 10.1016/j.jpba.2014.01.027

Goshtasby, A., Gage, S. H., and Bartholic, J. F. (1984). A two-stage cross correlation approach to template matching. *Pattern Anal. Mach. Intell. IEEE Trans.* 374–378. doi: 10.1109/TPAMI.1984.4767532

Gryniewicz-Ruzicka, C. M., Arzhantsev, S., Pelster, L. N., Westenberger, B. J., Buhse, L. F., and Kauffman, J. F. (2011). Multivariate calibration and instrument standardization for the rapid detection of diethylene glycol in glycerin by Raman spectroscopy. *Appl. Spectrosc.* 65, 334–341. doi: 10.1366/10-05976

Heraud, P., Wood, B. R., Beardall, J., and McNaughton, D. (2006). Effects of pre-processing of Raman spectra on *in vivo* classification of nutrient status of microalgal cells. *J. Chemometr.* 20, 193–197. doi: 10.1002/cem.990

Hutsebaut, D., Vandenabeele, P., and Moens, L. (2005). Evaluation of an accurate calibration and spectral standardization procedure for Raman spectroscopy. *Analyst* 130, 1204–1214. doi: 10.1039/b503624k

Juday, R. D. (1993). Optimal realizable filters and the minimum Euclidean distance principle. *Appl. Opt.* 32, 5100–5111. doi: 10.1364/AO.32.005100

Kompany-Zareh, M., and van den Berg, F. (2010). Multi-way based calibration transfer between two Raman spectrometers. *Analyst* 135, 1382–1388. doi: 10.1039/b927501k

Lee Rodgers, J., and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *Am. Stat.* 42, 59–66. doi: 10.2307/2685263

Li, Z., Wang, J. J., Huang, J., Zhang, Z. M., Lu, H. M., Zheng, Y. B., et al. (2013a). Nonlinear alignment of chromatograms by means of moving window fast Fourier transfrom cross-correlation. *J. Separat. Sci.* 36, 1677–1684. doi: 10.1002/jssc.201201021

Li, Z., Zhan, D. J., Wang, J. J., Huang, J., Xu, Q. S., Zhang, Z. M., et al. (2013b). Morphological weighted penalized least squares for background correction. *Analyst* 138, 4483–4492. doi: 10.1039/c3an00743j

Loethen, Y. L., Kauffman, J. F., Buhse, L. F., and Rodriguez, J. D. (2015). Rapid screening of anti-infective drug products for counterfeits using Raman spectral library-based correlation methods. *Analyst* 140, 7225. doi: 10.1039/C5AN01679G

Mann, C. K., and Vickers, T. J. (1999). Instrument-to-instrument transfer of Raman spectra. *Appl. Spectrosc.* 53, 856–861. doi: 10.1366/0003702991947441

McCreery, R. L. (2005). *Raman Spectroscopy for Chemical Analysis*. New York, NY: John Wiley & Sons.

Mrozek, M. F., Zhang, D., and Ben-Amotz, D. (2004). Oligosaccharide identification and mixture quantification using Raman spectroscopy and chemometric analysis. *Carbohydr. Res.* 339, 141–145. doi: 10.1016/j.carres.2003.09.019

Muehlethaler, C., Massonnet, G., and Esseiva, P. (2011). The application of chemometrics on infrared and raman spectra as a tool for the forensic analysis of paints. *Forensic Sci. Int.* 209, 173. doi: 10.1016/j.forsciint.2011.01.025

Nguyen Quang, H., Jouan, M., and Quy Dao, N. (1999). A simplified calibration model of spectral data for quantitative analyses with different Raman spectrometers. *Anal. Chim. Acta* 379, 159–167. doi: 10.1016/S0003-2670(98)00646-1

Papoulis, A. (1962). *The {F} ourier Integral and its Applications*. New York, NY.

Rodriguez, J. D., Gryniewicz-Ruzicka, C. M., Kauffman, J., Arzhantsev, S., Saettele, A. L., Berry, K. A., et al. (2013). Transferring raman spectral libraries and chemometric-based methods between different instruments and platforms. *Am. Pharm. Rev.* 16.

Rodriguez, J. D., Westenberger, B. J., Buhse, L. F., and Kauffman, J. F. (2011a). Quantitative evaluation of the sensitivity of library-based Raman spectral correlation methods. *Anal. Chem.* 83, 4061–4067. doi: 10.1021/ac200040b

Rodriguez, J. D., Westenberger, B. J., Buhse, L. F., and Kauffman, J. F. (2011b). Standardization of Raman spectra for transfer of spectral libraries across different instruments. *Analyst* 136, 4232–4240. doi: 10.1039/c1an15636e

Sacré, P.-Y., Deconinck, E., De Beer, T., Courselle, P., Vancauwenberghe, R., Chiap, P., et al. (2010). Comparison and combination of spectroscopic techniques for the detection of counterfeit medicines. *J. Pharm. Biomed. Anal.* 53, 445–453. doi: 10.1016/j.jpba.2010.05.012

Savitzky, A., and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36, 1627–1639. doi: 10.1021/ac60214a047

Taleb, A., Diamond, J., McGarvey, J. J., Beattie, J. R., Toland, C., and Hamilton, P. W. (2006). Raman microscopy for the chemometric analysis of tumor cells. *J. Phys. Chem. B* 110, 19625–19631. doi: 10.1021/jp061981q

Vickers, T. J., and Mann, C. K. (1999). Raman shift calibration of a compact multichannel spectrometer. *Appl. Spectrosc.* 53, 1617–1622. doi: 10.1366/0003702991946082

Wang, Y., Lysaght, M. J., and Kowalski, B. R. (1992). Improvement of multivariate calibration through instrument standardization. *Anal. Chem.* 64, 562–564. doi: 10.1021/ac00029a021

Wang, Y., Veltkamp, D. J., and Kowalski, B. R. (1991). Multivariate instrument standardization. *Anal. Chem.* 63, 2750–2756. doi: 10.1021/ac00023a016

Weatherall, J. C., Barber, J., Brauer, C. S., Johnson, T. J., Su, Y. F., Ball, C. D., et al. (2013). Adapting raman spectra from laboratory spectrometers to portable detection libraries. *Appl. Spectrosc.* 67, 149–157. doi: 10.1366/12-06759

Welch, L. (1974). Lower bounds on the maximum cross correlation of signals (Corresp.). *Inf. Theor. IEEE Trans.* 20, 397–399. doi: 10.1109/TIT.1974.1055219

Westad, F., and Martens, H. (1999). Shift and intensity modeling in spectroscopy—general concept and applications. *Chemometr. Intell. Lab. Syst.* 45, 361–370. doi: 10.1016/S0169-7439(98)00144-0

Wong, J. W., Cagney, G., and Cartwright, H. M. (2005a). SpecAlign—processing and alignment of mass spectra datasets. *Bioinformatics* 21, 2088–2090. doi: 10.1093/bioinformatics/bti300

Wong, J. W., Durante, C., and Cartwright, H. M. (2005b). Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.* 77, 5655–5661. doi: 10.1021/ac050619p

Zhang, Z. M., Chen, S., and Liang, Y. Z. (2010a). Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* 135, 1138–1146. doi: 10.1039/b922045c

Zhang, Z. M., Chen, S., Liang, Y. Z., Liu, Z. X., Zhang, Q. M., Ding, L. X., et al. (2010b). An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *J. Raman Spectrosc.* 41, 659–669. doi: 10.1002/jrs.2500

Zhang, Z. M., Liang, Y. Z., Lu, H. M., Tan, B. B., Xu, X. N., and Ferro, M. (2012). Multiscale peak alignment for chromatographic datasets. *J. Chromatogr. A* 1223, 93–106. doi: 10.1016/j.chroma.2011.12.047

Zou, M. Q., Zhang, X. F., Qi, X. H., Ma, H. L., Dong, Y., Liu, C. W., et al. (2009). Rapid authentication of olive oil adulteration by Raman spectrometry. *J. Agric. Food Chem.* 57, 6001–6006. doi: 10.1021/jf900217s

**Conflict of Interest Statement:** HC is employed by the company Shanghai Diracarta Biomedical Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

REVIEW
published: 26 October 2018
doi: 10.3389/fchem.2018.00503

# Wavelet Transform-Based UV Spectroscopy for Pharmaceutical Analysis

*Erdal Dinç*[1]* *and Zehra Yazan*[2]

[1] Department of Analytical Chemistry, Faculty of Pharmacy, Ankara University, Ankara, Turkey, [2] Department of Chemistry, Ankara University Faculty of Science, Ankara, Turkey

In research and development laboratories, chemical or pharmaceutical analysis has been carried out by evaluating sample signals obtained from instruments. However, the qualitative and quantitative determination based on raw signals may not be always possible due to sample complexity. In such cases, there is a need for powerful signal processing methodologies that can effectively process raw signals to get correct results. Wavelet transform is one of the most indispensable and popular signal processing methods currently used for noise removal, background correction, differentiation, data smoothing and filtering, data compression and separation of overlapping signals etc. This review article describes the theoretical aspects of wavelet transform (i.e., discrete, continuous and fractional) and its characteristic applications in UV spectroscopic analysis of pharmaceuticals.

**Keywords: discrete wavelet transform, continuous wavelet transform, fractional wavelet transform, UV spectroscopy, pharmaceutical analysis**

OPEN ACCESS

***Edited by:***
*Hoang Vu Dang,*
*Hanoi University of Pharmacy, Vietnam*

***Reviewed by:***
*Gaetano Ragno,*
*Università della Calabria, Italy*
*Joseph Dubrovkin,*
*Western Galilee College, Israel*

***Correspondence:**
*Erdal Dinç*
*dinc@ankara.edu.tr*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *27 April 2018*
***Accepted:*** *03 October 2018*
***Published:*** *26 October 2018*

***Citation:***
*Dinç E and Yazan Z (2018) Wavelet Transform-Based UV Spectroscopy for Pharmaceutical Analysis. Front. Chem. 6:503. doi: 10.3389/fchem.2018.00503*

## INTRODUCTION

In experimental studies, instruments or devices can provide signals (or graphs) in different formats e.g., spectrum, chromatogram, voltammogram, and electroferogram etc. The analysis of chemicals and pharmaceuticals in various samples is based upon the utilization of the measured signals of substances of interest. In practice, such an analysis for a multicomponent mixture may not be determined without a prior separation step due to spectral overlapping. Therefore, high performance liquid chromatography (HPLC) is one of the most commonly used techniques for quantitative estimation in the quality control of raw materials and commercial products in laboratories. In some cases, chromatographic determination could not be possible due to not only similar physicochemical behavior of analytes but also time- and solvent-consumption for optimal experimental conditions.

In practice, UV spectroscopic methods are widely used in chemical and pharmaceutical analysis. As compared to chromatographic ones, the use of spectroscopic methods provides a rapid analysis with low-cost and acceptable results. However, multicomponent analysis may not be possible with a traditional UV spectrophotometric approach due to spectral interferences of both active and inactive ingredients in samples. In some cases, derivative spectrophotometry (O'Haver and Green, 1976; O'Haver, 1979; Levillain and Fompeydie, 1986; Ragno et al., 2006) and its improved versions e.g., ratio spectra-derivative spectrophotometry (Salinas et al., 1990), ratio spectra-derivative spectrophotometry-zero crossing (Berzas Nevado et al., 1992; Dinç and Onur, 1998; Dinç, 1999), and double-divisior-ratio spectra-derivative spectrophotometry (Dinç and Onur, 1998; Dinç, 1999; Gohel et al., 2014; Shokry et al., 2014) could be used in place of

conventional UV spectrophotometric method for analysis of binary and ternary mixtures without using a separation step. However, these spectral approaches may not always yield successful data due to severely overlapping spectral bands, spectral noise and baseline variation. Additionally, high-order differentiation of spectra may lead to spectral deterioration i.e., a decrease in signal intensity and signal-to-noise ratio. As a result, a number of mathematical manipulations (or signal processing methods) are often required to make instrumental signals more meaningful for analysis purpose.

Generally speaking, transform (i.e., Fourier, Hilbert, short-time Fourier, Wigner distribution, Radon, and wavelet) is a very suitable technique in the pre-treatment step to simplify signals. Fourier transform (FT) is the first method to modify chemical signal (Griffiths, 1977; Cooper, 1978; Griffiths and De Haseth, 1986; Ernst, 1989) with the mathematical essence such as filtering, convolution/deconvolution etc. FT analysis can localize signal in frequency domain very well, but not so much in time domain. In contrast, wavelet transform (WT) has the advantage of localizing signals both in time (position) and frequency (scale) domains, making it a preferable mathematical tool to replace FT in the study of the local property of a signal and the removal of the perturbation of measuring error in spectral analysis. Nowadays, WT is one of the most signal analysis algorithms commonly used in the different fields of chemistry and engineering, providing alternative ways or opportunities to resolve complex spectral bands or diverse data types of signals.

For readers interested in learning the general theory of wavelets, more details can be found in the literature (Mallat, 1988; Chui, 1992; Daubechies, 1992; Newland, 1993; Byrnes et al., 1994; Chui et al., 1994; Vetterli and Kovačević, 1995; Strang and Nguyen, 1996).

In the signal smoothing and de-noising of spectral peaks, the elimination of noise requires an application of appropriate filters to the raw spectral data such as some conventional signal filters Savitzky–Golay, Fourier and Kalman (Brown et al., 1994, 1996). The use of WT in signal analysis is two-fold: (i) to detect the singularities of a signal very likely caused by high-frequency noise and (ii) to separate the signal frequencies at different scales (Palavajjhala et al., 1994; Yan-Fang, 2013; Li and Chen, 2014). To illustrate this, Barclay et al. (1997) performed a comparative study in de-noising and smoothing of Gaussian peak by using wavelet, Fourier and Savitzky–Golay filters i.e., smoothing eliminates high-frequency components of the transformed signal irrespective of their amplitudes, while de-noising eliminates small-amplitude components of the transformed signal irrespective of their frequencies.

Historically, WT principal applications in chemistry were first explored by Walczak and Massart (1997a), who presented an approach based on the application of wavelet packet transform (WPT) to the best-basis selection for the compression and de-noising of a set of signals in time-frequency domain. In their paper, the proposed technique was compared to Wickerhauser's approach (Wickerhauser, 1994) of fast approximate principal component analysis (PCA). These authors also published two more papers on the application of wavelets for data processing i.e., the introduction of WPT for noise suppression and signal compression (Walczak and Massart, 1997b) and the use of WT for signal compression and denoising, image processing, data compression and multivariate data modeling in analytical chemistry (Walczak and Massart, 1997c). On the other hand, Alsberg et al. (1997) tried to introduce WT to chemometricians by suggesting the short-time FT technique as a resolution to obtain information about frequency changes over time as well as the WT for de-noising, baseline removal, determination of derivative zero crossings and signal compression. In 1997, WT application in chemical analysis was also confirmed by Wang et al. (1997) and Depczynski et al. (1997). Up to date, WT processing of the different types of raw signals has been reported for liquid chromatography (Shao et al., 1997, 1998a,b,c) and NMR spectroscopy (Neue, 1996; Barache et al., 1997), Raman spectra (Cai et al., 2001; Ehrentreich and Summchen, 2001), and voltammetry (Chen et al., 1996; Fang and Chen, 1997; Zheng et al., 1998; Zhong et al., 1998; Aballe et al., 1999; Zheng and Mo, 1999) IR and Raman spectroscopy (Shao and Zhuang, 2004; Hwang et al., 2005; Chalus et al., 2007; Jun-fang et al., 2007; Lai et al., 2011). In this context, as in the various fields of mathematics and engineering, the implementations of WT in analytical chemistry and neighbor disciplines has become increasingly attractive as an alternative way to analyze complex mixtures previously unresolved by traditional analytical techniques.

With reference to the above-mentioned review, the aim of this paper is to describe the fundamentals of WT methodologies and its typical implementations for UV spectroscopic analysis of pharmaceuticals.

## BRIEF HISTORY OF WAVELETS

In the literature, the first study was related to the Haar Wavelet transform. This family was suggested by the mathematician Alfred Haar in 1909. However, the word "wavelet" was not used in the period of Haar. In fact, the word "wavelet" was invented by Morlet and the physicist Alex Grossman in 1984. After the first orthogonal Haar wavelet, the second orthogonal wavelet known as "Meyer wavelet" was formulated by the mathematician Yves Meyer in 1985. In 1988, Stephane Mallat and Meyer elaborated the concept of multiresolution. In the same year, a systematical method to construct compactly supported continuous wavelets was found by Ingrid Daubechies. Afterwards, Mallat proposed the fast wavelet transform. The emergence of this algorithm increased the implementations of the WT in the signal processing field.

In other words, the history of the wavelet families could be given in the following chronological order: Haar families in 1910, Morlet wavelet concept in 1981, Morlet and Grossman, "wavelet" in 1984, Meyer, "orthogonal wavelet" in 1985, Mallat and Meyer, multiresolution analysis in 1988, Daubechies, compact support orthogonal wavelet in 1988 and Mallat, fast wavelet transform in 1989 (c.f. Chun-Lin, 2010).

Basically, WT can be mainly classified into discrete wavelet transform (DWT) and continuous wavelet transform (CWT) in

the signal analysis. The theory and implementations of wavelets in chemistry and related fields were well documented as review papers (Leung et al., 1998; Dinç and Baleanu, 2007b; Dinç, 2013; Li and Chen, 2014; Medhat, 2015) and reference books (Walczak and Massart, 2000a,b; Walczak and Radomski, 2000; Brereton, 2003, 2008; Chau et al., 2004; Danzer, 2007; Mark and Workman, 2007; Dubrovkin, 2018).

## WAVELET TRANSFORM ALGORITHMS

FT is based upon the decomposition of a signal into a set of trigonometric (sine and cosine) functions i.e., FT represents a signal in terms of sinusoids. The representation of FT of a signal from time mode to frequency mode is illustrated in **Figure 1**. For the determination of a local information in the FT, it is required to use an analyzing function ψ having localization properties in both frequency and time domains. This $\psi$ function is named as a wavelet and it must be wave of finite duration.

WT contains the decomposition of a signal into a set of basic functions (wavelets). Basis functions of WT are small waves detected in different times. On the contrary to FT, WT gives information on both time and frequency, making it as an alternative approach to eliminate the resolution problem in signal analysis.

By definition, wavelets are the mathematical methods that convert the data into various coefficients and then analyze each coefficient at a resolution corresponding to its scale. Projection of a signal onto wavelet basic functions is called the wavelet transform. In other words, wavelets are mathematical functions generated from a mother wavelet Ψ(x) by the scaling parameter (dilatation) and shifting parameter (translation) i.e., the signal is expanded on a set of the dilatation (*scaling parameter*) of functions

$$\psi\left(\frac{x-a}{b}\right) \tag{1}$$

The scaling parameter has a significant role for the variation of time and frequency resolution when processing the signal.

For a given mother wavelet (Daubechies, 1992) $\psi(x)$ by the scaling parameter and shifting parameter o f$\psi(x)$, a set of functions expressed by $\psi_{a,b}(x)$ is obtained from the following equation.

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{x-b}{a}\right), \mathrm{a} \neq 0, \mathrm{a}, \mathrm{b} \in \mathrm{R} \tag{2}$$

where $a$ is the scaling parameter, $b$ is the shifting parameter and R is domain of real number. The mathematical expression of a CWT on a function $f$ (x) is given below

$$CWT\left\{f(x); a, b\right\} = \int_{-\infty}^{\infty} f(x)\psi^*_{a,b}(x)dx = \langle f(x), \psi_{a,b}\rangle \tag{3}$$

here the superscript * is related to the complex conjugate and $\langle f(x), \psi_{a,b}\rangle$ represents the inner product of function f(x) onto the wavelet function $\psi_{\mathrm{a,b}}$(x).

The original signal can be completely reconstructed by a sampled version of the CWT. Usually, the exemplar is follows as

$$a = 2^{-\mathrm{m}} \text{and } b = \mathrm{n}2^{-\mathrm{m}} \tag{4}$$

Here $a$ and $b$ denote scale and dilatation parameters, respectively, and $R$ is the real number. The expression of DWT can be given as

$$DWT = \int_{-\infty}^{+\infty} f(X)\ \psi^*_{m,n}(x)\,dt \tag{5}$$

Where $\psi^*_{\mathrm{m,n}}(X) = 2^{-\mathrm{m}}\ \psi\ (2^{\mathrm{m}}\ \mathrm{x} - \mathrm{n})$ is the dilated and translated version of the mother wavelet. In the application of the DWT, only outputs from the low-pass filter are processed by WT. However, in the wavelet packet decomposition of signals, both outputs from the low-pass and high-pass filters are manipulated by WT (Strang and Nguyen, 1996). Multiresolution decomposition with wavelets is an interesting topic for signal and image analysis (Mallat, 1988; Daubechies, 1992).

Some families of wavelets with names and their coding list are illustrated in **Table 1**.

For signal processing, there is also another WT approach i.e., fractional wavelet transform (FWT) specifically designed for rectification of the limitations of the WT and fractional FT (Blu and Unser, 2000, 2002; Unser and Blu, 2000). FWT is based on the fractional B-splines. As it is already known, the splines play an important role on the early development of the theory of WT.

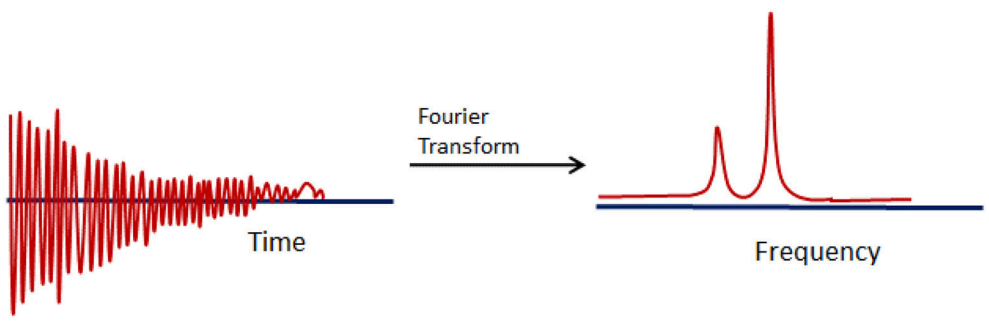



**FIGURE 1 |** Representation of Fourier transform of a signal from time domain to frequency domain.

**TABLE 1 |** Families of wavelets with names and their coding list.

| Wavelet families | Coding |
|---|---|
| Haar | haar |
| Daubechies | db |
| Symlets | sym |
| Coiflets | coif |
| BiorSplines | bior |
| ReverseBior | rbio |
| Meyer | meyr |
| Dmeyer | dmey |
| Gaussian | gaus |
| Mexican hat function | mexh |
| Morlet | morl |
| Complex Gaussian | cgau |
| Shannon | shan |
| Frequency B-Spline | fbsp |
| Complex Morlet | cmor |

A B-spline is generalization of the Beziers curve. Let a vector known as the knot be defined by $T = \{t_0, t_1, \ldots, t_m\}$ where T is a non-decreasing sequence with $t_i \epsilon [0, 1]$, and define control point $P_0$, $P_n$. The knots $t_0, t_1, \ldots, t_m$ is called internal knots. If $p = m - n - 1$ denotes the degree, the basis function is defined as follows:

$$N_{i,\,0}(t) = f\,(x) = \begin{cases} 1, \ if \ t_i \leq t \ < t_{i+1} \ and \ t_{i+1} \\ 0 \ otherwise \end{cases} \tag{6}$$

and

$$N_{i,\,p}\ (t) = \frac{t \ - \ t_i}{t_{i+p} \ - \ t_i} N_{i,\,p-1}\ (t) + \frac{t_{\,i+p+1} \ - \ t}{t_{i+p} + 1 \ - \ t_{i+1}} N_{i+1,\,p-1}\ (t) \tag{7}$$

Therefore, the curve defined by

$$C\,(t) = \sum\nolimits_{i=0}^{n} P_i\, N_{i,\,p}\,(t) \tag{8}$$

is a B-spline

Fractional B-spline: The fractional B-spline is defined as

$$\beta_+^{\alpha}\,(x) = \frac{\sum_{k=0}^{+\infty}\ (-1)^k \ \binom{\alpha+1}{k} \left(x-k\right)_+^{\alpha}}{\Gamma\ (\alpha+1)} \tag{9}$$

where Euler's Gamma function is obtained by

$$\Gamma\ (\alpha+1) = \int_0^{+\alpha} x^{\alpha}\ e^{-x}\ dx \tag{10}$$

and

$$(x-k)_{\alpha}^{+} = \max\,(x-k,\ 0)^{\alpha} \tag{11}$$

The forward fractional finite difference operator of order α is defined as

$$\Delta_+^{\alpha} f\,(x) = \sum\nolimits_{k=0}^{+\infty} (-1)^k\ \binom{\alpha}{k} f\ \left(x-k\right), \tag{12}$$

where

$$\binom{\alpha}{k} = \frac{\Gamma\ (\alpha+1)}{\Gamma\left(k+1\right)\left(\alpha-k+1\right)} \tag{13}$$

B-splines fulfill the convolution property, namely

$$\beta_+^{\alpha 1} * \beta_+^{\alpha 2} = \ \beta_+^{\alpha 1+\alpha 2} \tag{14}$$

The centered fractional B-splines of degree α is defined as

$$\beta_*^{\alpha}\,(x) = \ \frac{1}{\Gamma\ (\alpha+1)}\ \sum\nolimits_{k\epsilon Z} (-1)^k \left|\frac{\alpha+1}{k}\right|\ \left|x-k\right|_*^{\alpha} \tag{15}$$

where

$$|x|_-^{\alpha} = \begin{cases} \frac{|\ x|^{\alpha}}{-2\sin\left(\frac{\pi}{2}\alpha\right)}, \ \alpha \ not \ even \\ \frac{X\log x}{(-1)^{1+n}\,\pi}, \ \alpha \ even \end{cases} \tag{16}$$

The fractional B-spline wavelet is defined as

$$\psi_+^{\alpha}\ \left(\frac{x}{2}\right) = \sum\nolimits_{k\epsilon Z} \frac{(-1)^k}{2^{\alpha}} \sum\nolimits_{1\epsilon Z} \binom{\alpha+1}{1} \beta_*^{2\alpha+1}\left(1+k-1\right)\beta_+^{\alpha}\,(x-k) \tag{17}$$

We mention that the fractional splines wavelets of degree obey the following

$$\int_{-\infty}^{+\infty} X^n\ \psi_+^{\alpha}\ (x)\,dx = 0, \ \ldots, \ [\alpha] \tag{18}$$

and the Fourier transform fulfills the following relations

$$\hat{\psi}_+^{\alpha}\,(\varpi) = C\,(j\varpi)^{\alpha+1}, \ as\ \varpi \ \rightarrow 0 \tag{19}$$

and

$$\hat{\psi}_*^{\alpha}\,(\varpi) = C\,(j\varpi)^{\alpha+1}, \ as\ \varpi \ \rightarrow 0 \tag{20}$$

where $\hat{\psi}_+^{\alpha}\,(\varpi)$ is symmetric. The fractional spline wavelet behaves like a fractional derivative operator.

## STRATEGIES IN CWT APPLICATIONS TO UV SPECTROSCOPY ANALYSIS OF MULTICOMPONENT MIXTURES

For the past 15 years, the potential application of CWT in chemistry, especially in combination with other mathematical methods, leads us to a conclusion that WT has interestingly became a useful algorithm for UV quantitative analysis of pharmaceuticals. Four different models [i.e., continuous wavelet transform-zero crossing (CWT-ZC), ratio spectra-continuous wavelet transform (RS-CWT), ratio spectra-continuous wavelet transform-zero crossing (RS-CWT-ZC), and double divisor ratio spectra-continuous wavelet transform (DDRS-CWT)] were described in the implementation of CWT to UV spectroscopic data for the resolution of overlapping spectra to quantify drugs

in different types of samples. The modeling of CWT—UV spectroscopic approaches are detailed below. Fundamentally, these approached can be successfully applied to the UV spectroscopic analysis of binary and ternary mixtures, provided that the law of additivity of absorbance is obeyed.

## CONTINUOUS WAVELET TRANSFORM-ZERO CROSSING

The application of CWT-ZC approach to UV spectroscopic signals was first proposed by Dinç and Baleanu (2003a).

If a mixture of two analytes (M and N) is considered (see **Figure 2A**) and the absorbance of this binary mixture is measured at $\lambda_i$, we can have the following equation (Charlotte Grinter and Threlfall, 1992):

$$A_{mix,\,\lambda i} = \alpha_{M,\,\lambda i} C_M + \beta_{N,\,\lambda i} C_N \quad (21)$$

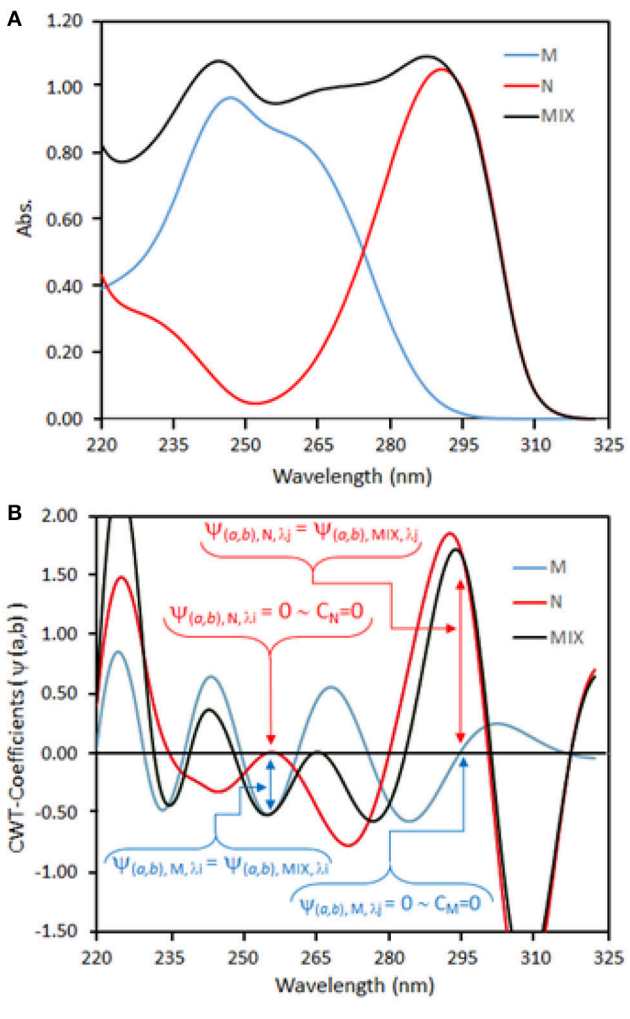



**FIGURE 2 | (A)** Absorption spectra and **(B)** CWT spectra of M (–) and N (–) compounds and their mixture (–).

where $Am_{\lambda i}$ is the absorbance of the binary mixture at wavelength $\lambda_i$, and the coefficients are the absorptivities of M and N, respectively. $C_M$ and $C_N$ represent the concentrations of M and N, respectively.

If CWT is applied to Equation (21), the following function can be obtained as

$$\psi_{(a.b),\,MIX,\,\lambda i} = \psi_{(a.b),M,\,\lambda i} C_M + \psi_{(a.b),\,N,\,\lambda i} C_N \quad (22)$$

If $\psi_{(a.b),N,\lambda i} C_N = 0$, then we obtain the following equation

$$\psi_{(a.b),\,MIX,\,\lambda i} = \psi_{(a.b),\,M,\,\lambda i} C_M \quad (23)$$

Equation (23) shows that CWT ($\psi_{(a.b),\,M,\,\lambda i}\, C_M$) amplitudes of M in the binary mixture are dependent only on $C_M$ regardless of $C_N$ (see **Figure 2B**).

## RATIO SPECTRA-CONTINUOUS WAVELET TRANSFORM

Apart from CWT-ZC approach, overlapping spectral bands in a binary mixture could be solved by the application of a combined hybrid approach i.e., RS-CWT (Dinç and Baleanu, 2004a,c).

The absorption spectra of M and N compounds, and their mixture are indicated in **Figure 3A**. By being divided by the standard spectrum ($A_{N,\lambda i} = \beta_{\lambda i}\, C_N^o$) of one of the compounds in the binary mixture, Equation (21) becomes

$$\frac{A_{m,\lambda i}}{\beta_{\lambda i}\, C_N{}^o} = \frac{\alpha_{\lambda i}\, C_M}{\beta_{\lambda i}\, C_N{}^o} + \frac{\beta_{\lambda i}\, C_N}{\beta_{\lambda i}\, C_N{}^o} \quad (24)$$

**Figure 3B** shows the ratio spectra of analytes and their binary mixture. If CWT is applied to Equation (24), the following equation can be obtained

$$CWT\left[\frac{A_{m,\lambda i}}{\beta_{\lambda i}\, C_N^o}\right] = CWT\left[\frac{\alpha_{\lambda i}}{\beta_{\lambda i}}\right]\frac{C_M}{C_N^o} + CWT\left[\frac{\beta_{\lambda i}}{\beta_{\lambda i}}\right]\frac{C_N}{C_N^o} \quad (25)$$

If $CWT\left[\frac{\beta_{\lambda i}}{\beta_{\lambda i}}\right]\frac{C_N}{C_N^o} = 0$, then we obtain

$$\mathrm{CWT}\left[\frac{A_{m,\lambda i}}{\beta_{\lambda i}\, C_N^o}\right] = \mathrm{CWT}\left[\frac{\alpha_{\lambda i}}{\beta_{\lambda i}}\right]\frac{C_M}{C_N^o} \quad (26)$$

The ratio-CWT amplitudes of the binary mixture given in Equation (26) depend only on $C_M$ and $C_N{}^o$ regardless of $C_N$ (e.g., see **Figure 3C**).

## RATIO SPECTRA-CONTINUOUS WAVELET TRANSFORM-ZERO CROSSING

In RS-CWT-ZC approach (Dinç et al., 2005a), if a mixture of three analytes (X, Y, and Z) is considered and the absorbance of this ternary mixture is measured at $\lambda_i$, the following mathematical expression (Charlotte Grinter and Threlfall, 1992) would be given

$$A_{mix,\,\lambda i} = \alpha_{X,\,\lambda i} C_X + \beta_{Y,\,\lambda i} C_Y + \gamma_{Z,\,\lambda i} C_Z \quad (27)$$

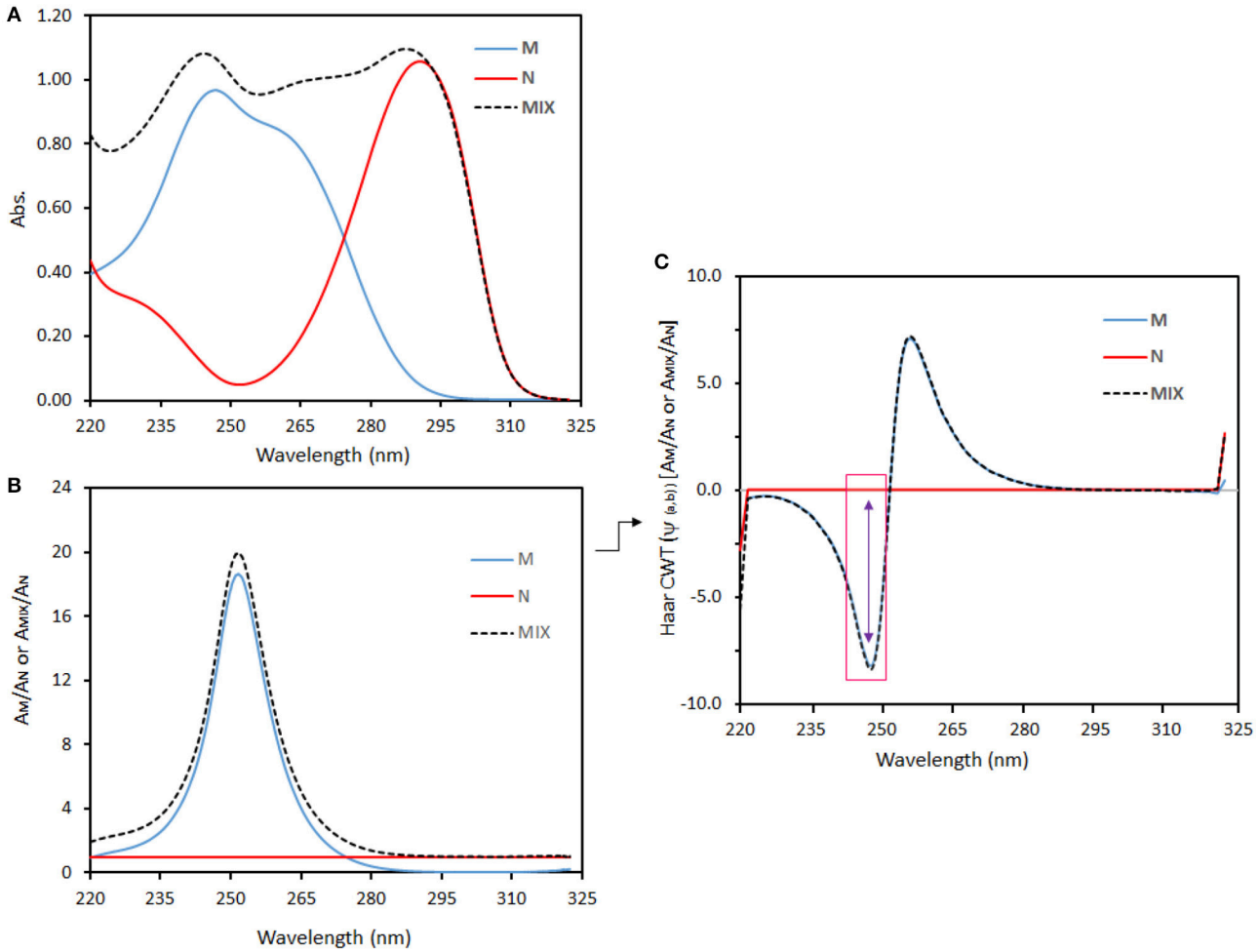



**FIGURE 3 | (A)** Absorption spectra, **(B)** ratio spectra, and **(C)** Haar CWT spectra of M and N compounds and their binary mixture.

Where $A_{mix,\,\lambda i}$ is the absorbance of the ternary mixture at wavelength $\lambda_i$, and coefficients $\alpha_{X,\,\lambda i}$, $\beta_{Y,\,\lambda i}$, and $\gamma_{Z,\,\lambda i}$ denote the absorptivities of X, Y, and Z, respectively. $C_X$, $C_Y$, and $C_Z$ represent the concentrations of X, Y, and Z, respectively.

If Equation (27) is divided by the spectrum of a standard solution ($C_X^o$) of one of the compounds in the ternary mixture, we have the following equation:

$$\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i}C_X{}^o} = \frac{\alpha_{X,\,\lambda i}C_X}{\alpha_{X,\,\lambda i}C_X{}^o} + \frac{\beta_{Y,\,\lambda i}C_Y}{\alpha_{X,\,\lambda i}C_X^o} + \frac{\gamma_{Z,\,\lambda i}C_Z}{\alpha_{X,\,\lambda i}C_X^o} \quad (28)$$

If CWT is applied to Equation (28), the following equation can be obtained

$$\mathrm{CWT}\left[\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i}C_X^o}\right] = \mathrm{CWT}\left[\frac{\beta_{Y,\,\lambda i}C_Y}{\alpha_{X,\,\lambda i}C_X^o}\right] + \mathrm{CWT}\left[\frac{\gamma_{Z,\,\lambda i}C_Z}{\alpha_{X,\,\lambda i}C_X^o}\right] \quad (29)$$

Equation (29) indicates that the CWT amplitudes of the ratio spectra of the ternary mixture are dependent only on $C_Z$ and $C_X{}^o$ regardless of the concentrations of other compounds.

## DOUBLE DIVISOR RATIO SPECTRA-CONTINUOUS WAVELET TRANSFORM

In addition to RS-CWT-ZC approach, the spectral resolution of ternary mixtures could be effectively done by DDRS-CWT approach (Dinç and Baleanu, 2008a) as follows.

When two compounds in the ternary mixture is used as a *double divisor*, we have

$$A^o_{mix,\,\lambda i} = \alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o \quad (30)$$

By dividing Equation (27) and (30), we obtain as follows

$$\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o} = \frac{\alpha_{X,\,\lambda i}C_X}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o} + \frac{\beta_{Y,\,\lambda i}C_Y}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o} + \frac{\gamma_{Z,\,\lambda i}C_Z}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o} \quad (31)$$

**TABLE 2 |** Applications of the continuous wavelet transform-zero crossing technique to UV spectroscopic analysis of pharmaceuticals.

| Pharmaceuticals | Method | Wavelet Families | Type of data | References |
|---|---|---|---|---|
| Thiamine HCl, pyridoxine HCl | CWT-zero crossing | Daubechies, Biorthogonal | UV absorption spectra | Dinç and Baleanu, 2003a |
| Hydrochlorothiazide, spironolactone | CWT-zero crossing | Daubechies, Biorthogonal | UV absorption spectra | Dinç et al., 2003 |
| Thiamine HCl; pyridoxine HCl | CWT-zero crossing | Mexican hat function, Meyer | UV absorption spectra | Dinç and Baleanu, 2003b |
| Thiamine HCl, pyridoxine HCl | CWT-zero crossing | Gaussian1, Gaussian2 | UV absorption spectra | Dinç and Baleanu, 2004a |
| Caffeine, propyphenazone | DWT-CWT-zero crossing | Mexican and Haar | UV absorption spectra | Dinç et al., 2004a |
| Benazepril, hydrochlorothiazide | DWT-CWT-zero crossing | Coiflets2 and Gaussian2 | UV absorption spectra | Dinç and Baleanu, 2004b |
| Hydrochlorothiazide, Spironolactone | CWT-zero crossing | Haar, Mexican hat function | UV absorption spectra | Dinç et al., 2004c |
| Benazepril, hydrochlorothiazide | CWT-zero crossing | Mexican, Haar, Daubechies3 | UV absorption spectra | Dinç and Baleanu, 2004c |
| Ascorbic acid, acetylsalicylic acid | CWT-zero crossing | Mexican hat function | UV absorption spectra | Dinç et al., 2005b |
| Diminazene aceturate and phenazone | CWT-zero crossing | Reverse Biorthogonal | UV absorption spectra | Dinç et al., 2005c |
| Quinapril, hydrochlorothiazide | CWT-zero crossing | Mexican hat wavelet function | UV absorption spectra | Dinç and Baleanu, 2007a |
| Oxfendazole and oxyclozanide | CWT-zero crossing | Mexican hat function | UV absorption spectra | Dinç and Baleanu, 2007c |
| Levodopa, benserazide | CWT-zero crossing | Symlets | UV absorption spectra | Dinç et al., 2007d |
| Chlortetracycline, benzocaine | CWT-zero crossing | Coiflets | UV absorption spectra | Dinç et al., 2007c |
| Pyridoxine hydrochloride, isoniazide | CWT-zero crossing | Mexican hat function | UV absorption spectra | Üstündag et al., 2008 |
| Risedronate sodium | CWT-zero crossing | Morlet, Biorthogonal | UV absorption spectra | Ugurlu et al., 2008 |
| ampicillin sodium, sulbactam sodium | CWT-zero crossing | Mexican hat function, Symtles | UV absorption spectra | Dinç and Baleanu, 2009a |
| Paracetamol, chloroxozone | CWT-zero crossing | Mexican hat function, Daubechies, Symplets, Coiflets, Biortogonal, Gaussian | UV absorption spectra | Dinç et al., 2009a |
| Levamisole, triclabendazole | CWT-zero crossing | Biorthogonal | UV absorption spectra | Dinç et al., 2009b |
| Telmisartan, hydrochlorothiazide | CWT-zero crossing | Gaussian, Biorthogonal | UV absorption spectra | Dinç and Baleanu, 2009b |
| Perindopril, indapamide | CWT-zero crossing | Haar and Biorthogonal1.5 | UV absorption spectra | Pektaş et al., 2009 |
| Valsartan, amlodipine | CWT-zero crossing | Daubechies, Dmeyer | UV absorption spectra | Dinç and Baleanu, 2010a |
| Metformin hydrochloride, glibenclamide | DWT-CWT-zero crossing | Daubechies, Reverse Biorthogonal, Gaussian | UV absorption spectra | Sohrabi et al., 2011 |
| Trimethoprim, sulphamethoxazole | CWT-zero crossing | Biorthogonal, Coiflets, Daubechies, Haar | UV absorption spectra | Dinç et al., 2011b |
| Amlodipine, atorvastatine | CWT-zero crossing | Mexican hat function | UV absorption spectra | Shariati-Rad et al., 2012 |
| Estradiol valerate, cyproterone acetate | CWT-zero crossing | Symlets | UV absorption spectra | Dinç et al., 2013a |
| Lamivudine, zidovudine | CWT-zero crossing | Mexican hat wavelet, Symlets, Daubechies | UV absorption spectra | Dinç et al., 2013b |
| Diphenhydramine hydrochloride | CWT-zero crossing | Biorthogonal | UV absorption spectra | Devrim et al., 2014 |
| Ambroxol hydrochloride, doxycycline | CWT-zero crossing | Haar wavelet function | UV absorption spectra | Darwish et al., 2014 |
| Oxfendazole, oxyclozanide | MOFrFT-CWT-zero-crossing | Mexican hat | UV absorption spectra | Dinç et al., 2015 |
| Atenolol, chlorthalidone | CWT-zero crossing | Coiflet, Mexican Hat function | UV absorption spectra | Dinç et al., 2017b |
| Valsartan, hydrochlorothiazide | CWT-zero crossing | Mexican hat function, Daubechies | UV absorption spectra | Dinç et al., 2017a |

Equation (31) can be simplified to

$$\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i}C_X{}^o + \beta_{Y,\,\lambda i}\,C_Y{}^o} = k + \frac{\gamma_{Z,\,\lambda i}C_Z}{\alpha_{X,\,\lambda i}C_X{}^o + \beta_{Y,\,\lambda i}\,C_Y{}^o} \quad (32)$$

Where $k = \frac{\alpha_{X,\,\lambda i}C_X + \beta_{Y,\,\lambda i}C_Y}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o}$ represents a constant for a given concentration range with respect to $\lambda_i$ in a certain region or point of wavelength.

A typical case is when $C_X{}^o$ and $C_Y{}^o$ are the same or very close to each other, namely $C_X{}^o = C_Y{}^o$ or $\cong C_X{}^o \cong C_Y{}^o$. Therefore, we obtain

$$\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o = C_X^o\,(\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}) \quad (33)$$

and Equation (32) can be written as

$$\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i}C_X^o + \beta_{Y,\,\lambda i}\,C_Y^o} = k + \frac{\gamma_{Z,\,\lambda i}C_Z}{C_X^o\,\left(\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}\right)} \quad (34)$$

After applying CWT to Equation (31), we have

$$CWT_{(a,b)}\left(\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}}\right)\frac{1}{C_X^o} = CWT_{(a,b)}\left(\frac{\gamma_{Z,\,\lambda i}C_Z}{\left(\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}\right)}\right)\frac{1}{C_X^o} \quad (35)$$

or

$$CWT_{(a,b)}\left(\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}}\right) = CWT_{(a,b)}\left(\frac{\gamma_{Z,\,\lambda i}}{\left(\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}\right)}\right)C_Z \quad (36)$$

In Equation (36), $C_Z$ is to proportional to the coefficients, $CWT_{(a,b)}\left(\frac{A_{mix,\,\lambda i}}{\alpha_{X,\,\lambda i} + \beta_{Y,\,\lambda i}}\right)$, at $\lambda i$. If this procedure is separately applied for pure Z and its ternary mixture, the $CWT_{(a,b)}$ coefficients are coincided at some characteristic point or region of wavelength, independent upon both $C_X$ and $C_Y$.

## WAVELET TRANSFORM-BASED UV SPECTROSCOPIC ANALYSIS OF PHARMACEUTICALS

Typical applications of CWT and FWT algorithms for UV spectroscopic analysis of pharmaceuticals are displayed in **Tables 2–5**. It is worth mentioning that WT could be solely applied to raw spectra and ratio spectra (as above-specified) as well as utilized as a hybrid approach (FWT-derivative, FWT-CWT-zero crossing, WT combined with multivariate calibration) for the simultaneous determination of analytes in pharmaceutical binary and ternary mixtures. It was shown that wavelet analysis of UV spectroscopic data was performed by using Wavelet Toolbox and m-file in MATLAB software. The numerous works provided by Dinç and co-workers have clearly highlighted the success of WT-based UV spectroscopic analysis for multicomponent synthetic mixtures, veterinary and pharmaceutical dosage forms as well as different types of test (e.g., assay, *in vitro* dissolution, stability indicating). Most studies proved it to be suitable for the routine analysis of dosage forms with good precision and accuracy, comparable to HPLC.

## CONCLUSIONS

In the point of view of UV spectroscopic analysis of multicomponent mixtures, CWT-based UV spectroscopic

**TABLE 3 |** Applications of the wavelet transform-multivariate approaches to UV spectroscopic analysis of pharmaceuticals.

| Pharmaceuticals | Method | Families | Type of data | References |
|---|---|---|---|---|
| Tetramethrin, propoxur; piperonil butoxide | CWT-PCR, CWT-PLS | Mexican hat function | UV absorption spectra | Dinç et al., 2004b |
| Paracetamol, ascorbic acid, acetylsalicylic acid | DWT-CLS, DWT-PLS | Haar | UV absorption spectra | Dinç et al., 2006a |

**TABLE 4 |** Applications of the ratio spectra-continuous wavelet transform, ratio spectra- continuous wavelet transform-zero crossing approaches to UV spectroscopic analysis of pharmaceuticals.

| Pharmaceuticals | Method | Families | Type of data | References |
|---|---|---|---|---|
| Paracetamol, acetylsalicylic acid, caffeine | Ratio spectra-CWT-ZC | Mexican hat function | UV ratio spectra | Dinç et al., 2005a |
| Diminazene aceturate and phenazone | Ratio spectra-CWT | Reverse Biorthogonal | UV ratio spectra | Dinç et al., 2005c |
| Paracetamol, metamizol, caffeine | Ratio spectra-CWT-ZC | Mexican hat function, Reverse biorthogonal, Biorthogonal | UV ratio spectra | Dinç et al., 2006b |
| Levamizol, oxycloanide | Ratio spectra-CWT | Daubechies | UV ratio spectra | Dinç et al., 2007a |
| oxfendazole and oxyclozanide | Ratio spectra-CWT | Morlet | UV ratio spectra | Dinç and Baleanu, 2007c |
| Ascorbic acid, acetylsalicylic acid and paracetamol | Double divisor-ratio spectra-CWT | Haar, Mexican hat function | UV- double divisor-ratio spectra | Dinç and Baleanu, 2008a |
| vitamin C, aspirin | Ratio spectra-CWT | Biorthogonal | UV ratio spectra | Dinç and Baleanu, 2008b |
| valsartan and hydrochlorothiazide | Ratio spectra-CWT | Mexican hat function, Coiflets | UV ratio spectra | Dinç et al., 2017a |

**TABLE 5 |** Applications of the fractional wavelet transform and its combination with other chemometric techniques to UV spectroscopic analysis of pharmaceuticals.

| Pharmaceuticals | Method | Families | Type of data | References |
|---|---|---|---|---|
| Ampicillin, sulbactam | FWT-derivative method | – | UV absorption data | Dinç and Baleanu, 2006 |
| Lacidipine and its photodegradation product | FWT-CWT | Mexican hat function | UV absorption data | Dinç et al., 2006c |
| Cilazapril, hydrochlorothiazide | FWT-PLS | – | UV absorption data | Dinç et al., 2007b |
| Paracetamol, propiphenazone, caffeine and thiamine | FWT-PCR, FWT-PLS, FWT-ANN | – | UV absorption data | Dinç et al., 2008 |
| Amlodipine, valsartan | FWT-PLS1, FWT-PLS2 | – | UV absorption data | Çelebier et al., 2010 |
| Trimethoprim, sulfachloropyridazine sodium | FWT-derivative method | – | UV absorption data | Kanbur et al., 2010 |
| Atorvastatin, amlodipine | FWT-CWT | Mexican wavelet hat function | UV absorption data | Dinç and Baleanu, 2010b |
| Trimethoprim, sulphamethoxazole | FWT-PCR, FWT-PLS | – | UV absorption data | Dinç et al., 2010 |
| of oxytetracycline and flunixin megluminein | FWT-PCR, FWT-PLS | – | UV absorption data | Kambur et al., 2011 |
| Olmesartan modoxomil, hydrochlorothiazide | FWT-CWT | Mexican wavelet hat function | UV absorption data | Dinç et al., 2011a |
| Thiamine HCl, pyridoxine HCl, lidocaine HCl | FWT-PCR, FWT-PLS, FWT-CWT-PCR, FWT-CWT-PLS | – | UV absorption data | Dinç and Baleanu, 2012 |
| Melatonin and its photodegradation | FWT-CWT | Biorthogonal, symplets | UV absorption data | Dinç et al., 2012 |

methods have outperformed both conventional and derivative UV spectroscopy in resolving spectrally binary and ternary mixtures. Nevertheless, wavelet analysis may not also have a sufficient power to resolve overlapping spectra of analytes in samples due to similarity of molecular structures and signal frequencies in some cases. They may not give desirable results for a complex mixture containing more than three compounds and/or a significant difference in ratios of active ingredients. In such a case, the use of WT coupled with chemometric PLS and PCR calibrations is advisable. Undoubtedly, however, wavelets can still be used as a mathematical prism for signal analysis because they can offer many possibilities such as baseline correction, noise removal and resolution of overlapping peaks, when the frequencies of analyzed components are significantly different from each other.

## AUTHOR CONTRIBUTIONS

Contributions of ED are planning and writing of the review paper. Contributions of ZY are literature review, collection, editing, and format arrangement.

## REFERENCES


Aballe, A., Bethencourt, M., Botana, F. J., and Marcos, M. (1999). Using wavelet transform in the analysis of electrochemical noise data. *Electrochim. Acta* 44, 4805–4816. doi: 10.1016/S0013-4686(99)00222-4

Alsberg, B. K., Woodward, A. M., and Kell, D. B. (1997). An introduction to wavelet transform for chemometricians: a time-frequency approach. *Chemom. Intell. Lab. Syst.* 37, 215–239. doi: 10.1016/S0169-7439(97)00029-4

Barache, D., Antoine, J. P., and Dereppe, J. M. (1997). The continuous wavelet transform, an analysis tool for NMR spectroscopy. *J. Magn. Reson.* 128, 1–11. doi: 10.1006/jmre.1997.1214

Barclay, V. J., Bonner, R. F., and Hamilton, I. P. (1997). Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Anal. Chem.* 69, 78–90. doi: 10.1021/ac960638m

Berzas Nevado, J. J., Guiberteau Cabanillas, C., and Salinas, F. (1992). Spectrophotometric resolution of ternary mixtures of salicylaldehyde, 3-hydroxybenzaldehyde and 4-hydroxybenzaldehyde by the derivative ratio spectrum-zero crossing method. *Talanta* 39, 547–553. doi: 10.1016/0039-9140(92)80179-H

Blu, T., and Unser, M. (2000). "The fractional spline wavelet transform: definition and implementation," in *Proceedings of the Twenty-Fifth IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00)* (Istanbul), 512–515.

Blu, T., and Unser, M. (2002). Wavelets, fractals, and radial basis functions. *IEEE Trans. Signal Process.* 50, 543–553. doi: 10.1109/78.984733

Brereton, R. G. (2003). *Data Analysis for the Laboratory and Chemical Plant*. New York, NY: Wiley & Sons, Inc., 167.

Brereton, R. G. (2008). *Applied Chemometrics for Scientists*. John-Wiley & sons Ltd.

Brown, S., Blank, T. B., Sum, S. T., and Weyer, L. G. (1994). Chemometrics. *Anal. Chem.* 66, 315R−359R. doi: 10.1021/ac00084a014

Brown, S., Sum, S. T., and Despagne, F. (1996). Chemometrics. *Anal. Chem.* 68, 21R−62R. doi: 10.1021/a1960005x

Byrnes, J. S., Byrnes, J. L., Hargreaves, K. A., and Berry, K. (1994). *Wavelet and Their Application*. Netherlands: Kluwer Academic Publishers.

Cai, W. S., Wang, L. Y., Pan, Z. X., Zuo, J., Xu, C. Y., and Shao, X. G. (2001). Application of the wavelet transform method in quantitative analysis of raman spectra. *J. Raman Spectrosc.* 32, 207–209. doi: 10.1002/jrs.688

Çelebier, M., Altinöz, S., and Dinç, E. (2010). "Fractional wavelet transform and chemometric calibrations for the simultaneous determination of amlodipine and valsartan in their complex mixture," in *New Trends in Nanotechnology and Fractional Calculus*, eds D. Baleanu, Z. B. Güvenç, and J. A. Tenreiro Machado (Dordrecht; Heidelberg; London; New York, NY: Springer), 333–340.

Chalus, P., Walter, S., and Ulmschneider, M. (2007). Combined wavelet transform–artificial neural network use in tablet active content determination by near-infrared spectroscopy. *Anal. Chim. Acta* 591, 219–224. doi: 10.1016/j.aca.2007.03.076

Charlotte Grinter, H., and Threlfall, T. L. (1992). *UV-VIS Spectroscopy and its Applications*. Berlin; Heidelberg; New York, NY: Springer-Verlag.

Chau, F., Yi-Zeng, L., Junbin, G., Xue-Guang, S., and James, D., W. (2004). Chemometrics, from Basics to *Wavelet Transform 1st Edn*. Hoboken, NJ: Wiley-Interscience.

Chen, J., Zhong, H. B., Pan, Z. X., and Zhang, M. S. (1996). Application of the wavelet transform in differential pulse voltammetric data processing. *Chin. J. Anal. Chem.* 24, 1002–1006.

Chui, C. K. (1992). *An Introduction to Wavelets*. New York, NY: Academic Press. 49.

Chui, C. K., Montefusco, L., and Puccio, L. (1994). *Wavelets: Theory, Algorithms and Applications*. San Diego, CA: Academic Press.

Chun-Lin, L. (2010). *A Tutorial of the Wavelet Transform*. Available online at: http://disp.ee.ntu.edu.tw/tutorial/WaveletTutorial.pdf

Cooper, J. W. (1978). "Chapter: Data handling in fourier transform spectroscopy," in *Transform Techniques in Chemistry,* ed P. R. Griffiths (New York, NY: Plenum Press), 4, 69–108.

Danzer, K. (2007). *Analytical Chemistry. Theoretical and Metrological Fundamentals*. Berlin; New York, NY: Springer.

Darwish, H. W., Metwally, F. H., and El.Bayoumi, A. (2014). Application of continous wavelet transform for derivative spectrophotometric determination of binary mixture in pharmaceutical dosage form. *Dig. J. Nanomater. Biostruct.* 9, 7–18.

Daubechies, I. (1992). *Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics*. Philadelphia, PA.

Depczynski, U., Jetter, K., Molt, K., and Niemoller, A. (1997). The fast wavelet transform on compact intervals as a tool in chemometrics. I. Mathematical background. *Chem. Intell. Lab. Syst*. 39, 19–27. doi: 10.1016/S0169-7439(97)00068-3

Devrim, B., Dinç, E., and Bozkir, A. (2014). Fast determination of diphenhydramine hydrochloride in reconstituted table syrups by CWT, PLS and PCR methods. *Acta Pol. Pharm*. 71, 721–729.

Dinç, E. (1999). The spectrophotometric multicomponent analysis of a ternary mixture of ascorbic acid, acetylsalicylic acid and paracetamol by the double divisor-ratio spectra derivative and ratio spectra-zero crossing methods. *Talanta* 48, 1145–1157. doi: 10.1016/S0039-9140(98)00337-3

Dinç, E. (2013). Wavelet transforms and applications in drug analysis, *FABAD J. Pharm. Sci.* 38, 159–165.

Dinç, E., Arslan, F., and Baleanu, D. (2009a). Alternative approaches to the spectral quantitative resolution of two-component mixture by wavelet families. *J. Chil. Chem. Soc*. 54, 28–35. doi: 10.4067/S0717-97072009000100007

Dinç, E., and Baleanu, D. (2003a). Multidetermination of thiamine HCl and pyridoxine HCl in their mixture using continuous daubechies and biorthogonal wavelet analysis. *Talanta* 59, 707–717. doi: 10.1016/S0039-9140(02)00611-2

Dinç, E., and Baleanu, D. (2003b). A zero-crossing technique for the multidetermination of thiamine HCl and pyridoxine HCl in their mixture by using one-dimensional wavelet transform. *J. Pharm. Biomed. Anal*. 31, 969–978. doi: 10.1016/S0731-7085(02)00705-7

Dinç, E., and Baleanu, D. (2004a). Multicomponent quantitative resolution of binary mixtures using continuous wavelet transform. *J. AOAC Int.* 87, 360–365.

Dinç, E., and Baleanu, D. (2004b). Application of the wavelet method for the simultaneous quantitative determination of benazepril and hydrochlorothiazide in their mixtures. *J. AOAC Int*. 87, 834–841.

Dinç, E., and Baleanu, D. (2004c). One-dimension continuous wavelet resolution for the simultaneous analysis of binary mixture of benazepril and hydrochlorothiazide in tablets using spectrophotometric absorbance data. *Rev. Roum. Chim*. 49, 917–925.

Dinç, E., and Baleanu, D. (2006). A new fractional wavelet approach for the simultaneous determination of ampicillin sodium and sulbactam sodium in a binary mixture. *Spectrochim. Acta Part A* 63, 631–638. doi: 10.1016/j.saa.2005.06.012

Dinç, E., and Baleanu, D. (2007a). Continuous wavelet transform and chemometric methods for quantitative resolution of a binary mixture of quinapril and hydrochlorothiazide in tablets. *J. Braz. Chem. Soc.* 18, 962–968. doi: 10.1590/S0103-50532007000500013

Dinç, E., and Baleanu, D. (2007b). "A review on the wavelet transforms applications in analytical chemistry," in *Mathematical Methods in Engineering*, eds K. Taş, J. A. Tenreiro Machado, and D. Balanu (Dordrecht: Springer), 265–285.

Dinç, E., and Baleanu, D. (2007c). Continuous wavelet transform applied to the overlapping absorption signals and their ratio signals for the quantitative resolution of mixture of oxfendazole and oxyclozanide in bolus. *J. Food Drug Anal*. 15, 109–117.

Dinç, E., and Baleanu, D. (2008a). Application of haar and mexican hat wavelets to double divisor-ratio Spectra for the multicomponent determination of ascorbic acid, acetylsalicylic acid and paracetamol in effervescent tablets. *J. Braz. Chem. Soc*. 19, 434–444. doi: 10.1590/S0103-50532008000300010

Dinç, E., and Baleanu, D. (2008b). Ratio spectra-continuous wavelet transform and ratio spectra-derivative spectrophotometry for the quantitative analysis of effervescent tablets of vitamin C and aspirin. *Rev. Chim. (Bucuresti)* 59, 499–504.

Dinç, E., and Baleanu, D. (2009a). Continuous wavelet transform applied to the quantitative analysis of a binary mixture. *Rev. Chim. (Bucuresti)* 60, 216–221.

Dinç, E., and Baleanu, D. (2009b). Spectral continuous wavelet transform for the simultaneous spectrophotometric analysis of a combined pharmaceutical formulation. *Rev. Chim.* 60, 741–744.

Dinç, E., and Baleanu, D. (2010a). Continuous wavelet transform applied to the simultaneous spectrophotometric determination of valsartan and amlodipine in tablets. *Rev. Chim.* 61, 290–294.

Dinç, E., and Baleanu, D. (2010b). Fractional wavelet transform for the quantitative spectral resolution of the composite signals of the active compounds in a two-component mixture. *Comput. Math. Appl.* 59, 1701–1708. doi: 10.1016/j.camwa.2009.08.012

Dinç, E., and Baleanu, D. (2012). Fractional-continuous wavelet transforms and ultra-performance liquid chromatography for the multicomponent analysis of a ternary mixture containing thiamine, pyridoxine, and lidocaine in ampules. *J. AOAC Int*. 95, 903–912. doi: 10.5740/jaoacint.11-199

Dinç, E., Baleanu, D., and Abaul-Enein, H. Y. (2004a). Wavelet analysis for the multicomponent determination in a binary mixture of caffeine and propyphenazone in tablets. *IL Farmaco* 59, 335–342. doi: 10.1016/j.farmac.2004.01.002

Dinç, E., Baleanu, D., Ioele, G., De Luca, M., and Ragno, G. (2008). Multivariate analysis of paracetamol, propyphenazone, caffeine and thiamine in quaternary mixtures by PCR, PLS and ANN calibrations applied on wavelet transform data. *J. Pharm. Biomed. Anal*. 48, 1471–1475. doi: 10.1016/j.jpba.2008.09.035

Dinç, E., Baleanu, D., and Kanbur, M. (2005c). A comparative application of wavelet approaches to the absorption and ratio spectra for the simultaneous determination of diminazene aceturate and phenazone in veterinary granules for injection. *Pharmazie* 60, 892–896.

Dinç, E., Baleanu, D., and Taş, A. (2006b). Wavelet transforms and artificial neural network for the quantitative resolution of ternary mixtures. *Rev. Chim.* 57, 626–631.

Dinç, E., Baleanu, D., and Taş, K. (2007a). "Continuous wavelet analysis for the ratio signals of the absorption spectra of binary mixtures," in *Mathematical Methods* in *Engineering,* ed Taş, K., J. A. Tenreiro Machado, and D. Baleanu (Dordrecht: Springer), 285–293.

Dinç, E., Baleanu, D., and Taş, K. (2007b). Fractional wavelet analysis for the composite signals of two-component mixture by multivariate spectral calibration. *J. Vib. Control* 13, 1283–1290. doi: 10.1177/1077546307077464

Dinç, E., Baleanu, D., Üstündag, Ö. (2003). An approach to quantitative two-component analysis of a mixture containing hydrochlorothiazide and spironolactone in tablets by one-dimensional continuous daubechies and biorthogonal wavelet analysis of UV-spectra, *Spectros. Lett*. 36, 341–355. doi: 10.1081/SL-120024583

Dinç, E., Baleanu, D., Üstündag, Ö., and Abaul-Enein, H. Y. (2004c). Continuous wavelet transformation applied to the simultaneous quantitative analysis of two–component mixtures. *Pharmazie* 59, 618–623.

Dinç, E., Büker, E., and Baleanu, D. (2011a). Fractional and continuous wavelet transforms for the simultaneous spectral analysis of a binary mixture system. *Commun. Nonlinear Sci. Numer. Simul.* 16, 4602–4609. doi: 10.1016/j.cnsns.2011.02.018

Dinç, E., Demirkaya, F., Baleanu, D., Kadioglu,Y., and Kadioglu, E. (2010). New approach for simultaneous spectral analysis of a complex mixture using the fractional wavelet transform. *Commun. Nonlinear Sci. Numer. Simul.* 15, 812–818. doi: 10.1016/j.cnsns.2009.05.021

Dinç, E., Duarte, F. B., Tenreiro Machado, J. A., and Baleanu, D. (2015). Application of continuous wavelet transform to the analysis of the modulus

of the fractional fourier transform bands for resolving two component mixture. *Signal Image Video P.* 9, 801–807. doi: 10.1007/s11760-013-0503-9

Dinç, E., Kadioglu, Y., Demirkaya, F., and Baleanu, D. (2011b). Continuous wavelet transforms for simultaneous spectral determination of trimethoprim and sulphamethoxazole in tablets. *J. Iran. Chem. Soc.* 8, 90–99. doi: 10.1007/BF03246205

Dinç, E., Kanbur, M., and Baleanu, D. (2007c). Comparative spectral analysis of veterinary powder product by continuous wavelet and derivative transforms. *Spectrochim. Acta Part A.* 68, 225–230. doi: 10.1016/j.saa.2006.11.018

Dinç, E., Kaş, F., and Baleanu, D. (2013a). A signal processing tool based on the continuous wavelet for the simultaneous determination of estradiol valerate and cyproterone acetate in their mixtures. *Rev. Chim.* 64, 124–126.

Dinç, E., Kaya, S., Doganay, D., and Baleanu, D. (2007d). Continuous wavelet and derivative transforms for the simultaneous quantitative analysis and dissolution test of levodopa-benserazide tablets. *J. Pharm. Biomed. Anal.* 44, 991–995. doi: 10.1016/j.jpba.2007.03.027

Dinç, E., and Onur, F. (1998). Application of a new spectrophotometric method for the analysis of a ternary mixture containing metamizol, paracetamol and caffeine in tablets. *Anal. Chim. Acta* 359, 93–106. doi: 10.1016/S0003-2670(97)00615-6

Dinç, E., Özdemir, A., and Baleanu, D. (2005a). An application of derivative and continuous wavelet transforms to the overlapping ratio spectra for the quantitative multiresolution of a ternary mixture of paracetamol, acetylsalicylic acid and caffeine in tablets. *Talanta* 65, 36–47. doi: 10.1016/j.talanta.2004.05.011

Dinç, E., Özdemir, A., and Baleanu, D. (2005b). Comparative study of the continuous wavelet transform, derivative and partial least squares methods applied to the overlapping spectra for the simultaneous quantitative resolution of ascorbic acid and acetylsalicylic acid in effervescent tablets. *J. Pharm. Biomed. Anal.* 37, 569–575. doi: 10.1016/j.jpba.2004.11.020

Dinç, E., Özdemir, A., Baleanu, D., and Taş, K. (2006a). Wavelet transform with chemometrics techniques for quantitative multiresolution analysis of a ternary mixture consisting of paracetamol, ascorbic acid and acetylsalicylic acid in effervescent tablets. *Rev. Chim. (Bucharest)* 57, 505–510.

Dinç, E., Özdemir, N., Üstündag, Ö., and Günseli Tilkan, M. (2013b). Continuous wavelet transforms for the simultaneous quantitative analysis and dissolution testing of lamivudine–zidovudine tablets. *Chem. Pharm. Bull.* 61, 1220–1227. doi: 10.1248/cpb.c13-00284

Dinç, E., Pektaş, G., and Baleanu, D. (2009b). Continuous wavelet transform and derivative spectrophotometry for the quantitative spectral resolution of a mixture containing levamizole and triclabendazole in veterinary tablets. *Rev. Anal. Chem.* 28, 79–92. doi: 10.1515/REVAC.2009.28.2.79

Dinç, E., Ragno, G., Baleanu, D., De Luca, M., and Ioele, G. (2012). Fractional wavelet transform-continuous wavelet transform for the quantification of melatonin and its photodegradation product. *Spectrosc. Lett.* 45, 337–343. doi: 10.1080/00387010.2012.666699

Dinç, E., Ragno, G., Ioele, G., and Baleanu, D. (2006c). Fractional wavelet analysis for the simultaneous quantitative analysis of lacidipine and its photodegradation product by continuous wavelet transform and multilinear regression calibration. *JAOAC Int.* 89, 1538–1546.

Dinç, E., Saygeçitli, E., and Ertekin, Z. C. (2017b). Simultaneous determination of atenolol and chlorthalidone in tablets by wavelet transform methods. *FABAD J. Pharm. Sci.* 42, 103–109.

Dinç, E., Üstündağ, Ö., Yüksel Tilkan, G., Türkmen, B., and Özdemir, N. (2017a). Continuous wavelet transform methods for the simultaneous determinations and dissolution profiles of valsartan and hydrochlorothiazide in tablets. *Braz. J. Pharm. Sci.* 53:e16050. doi: 10.1590/s2175-97902017000116050

Dinç, E., Baleanu, D., and Kanbur, M. (2004b). Spectrophotometric multicomponent determination of tetramethrin, propoxur and piperonyl butoxide in insecticide formulation by principal component regression and partial least squares techniques with continuous wavelet transform. *Can. J. Anal. Sci. Spect.* 49, 218–225.

Dubrovkin, J. (2018). *Mathematical Processing of Spectral Data in Analytical Chemistry: A Guide to Error Analysis*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Ehrentreich, F., and Summchen, L. (2001). Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal. Chem.* 73, 4364–4373. doi: 10.1021/ac0013756

Ernst, R. R. (1989). Nuclear magnetic resonance fourier transform spectroscopy. *Bull. Magn. Resonance* 16, 5–29.

Fang, H., and Chen, H. Y. (1997). Wavelet analyses of electroanalytical chemistry responses and an adaptive wavelet filter. *Anal. Chim. Acta* 346, 319-−325. doi: 10.1016/S0003-2670(97)90071-4

Gohel, R. V., Parmar, S. J., and Patel, B. A. (2014). Development and validation of double divisor-ratio spectra derivative spectrophotometric method for simultaneous estimation of olmesartan medoxomil, amlodipine besylate and hydrochlorthiazide in tablet dosage form. *Int. J. Pharm.Tech. Res.* 6, 1518–1525.

Griffiths, P. R. (1977). Recent applications of fourier transform infrared spectrometry in chemical and environmental analysis. *Appl. Spectros.* 31, 497–505. doi: 10.1366/000370277774464084

Griffiths, P. R., and De Haseth, J. A. (1986). *Fourier Transform Infrared Spectroscopy*. New York, NY: Wiley.

Hwang, M. S., Choa, C., Chunga, H., and Woo, Y. A. (2005). Nondestructive determination of the ambroxol content in tablets by Raman spectroscopy. *J. Pharm. Biomed. Anal.* 38, 210–215. doi: 10.1016/j.jpba.2004.12.031

Jun-fang, X., Xiao-yu, L., Pei-wu, L., Qian, M. A., and Xiao-xia, D. (2007). Application of wavelet transform in the prediction of navel orange vitamin C content by near-infrared spectroscopy. *Agr. Sci. China* 6, 1067–1073. doi: 10.1016/S1671-2927(07)60148-5

Kambur, M., Narin, I., Dinç, E., Candir, S., and Baleanu, D. (2011). Fractional wavelet transform for the quantitative spectral resolution of the commercial veterinary preparations. *Rev. Chim. (Bucuresti)* 62, 618–621.

Kanbur, M., Narin, I., Özdemir, E., and Dinç, E. (2010). "Fractional wavelet transform for the quantitative spectral analysis of two-component system," *New Trends in Nanotechnology and Fractional Calculus*, eds D. Baleanu, Z. B. Güvenç, and J. A. Tenreiro Machado (Dordrecht; Heidelberg; London; New York, NY: Springer), 321–331.

Lai, Y., Nia, N., and Kokot, S. (2011). Discrimination of Rhizoma Corydalis from two sources by near-infrared spectroscopy supported by the wavelet transform and least-squares support vector machine methods. *Vib. Spectrosc.* 56, 154–160. doi: 10.1016/j.vibspec.2011.01.007

Leung, A. K. M., Chau, F.-T., and Gau, J.-B. (1998). A review on applications of wavelets transform techniques in chemical analysis, 1989-1997. *Chem. Intell. Lab. Sys.* 43, 165–184. doi: 10.1016/S0169-7439(98)00080-X

Levillain, P., and Fompeydie, D. (1986). Spectrophotométrie dérivée: intérêt, limites et applications. *Analusis* 14, 1–20.

Li, B., and Chen, X. (2014). Wavelet-based numerical analysis: a review and classification. *Finite Elem. Anal. Des.*, 81, 14–31. doi: 10.1016/j.finel.2013.11.001

Mallat, S. (1988). *A Wavelet Tour of Signal Processing*. New York, NY: Academic Press.

Mark, H., and Workman, J. (2007). *Chemometrics in Spectroscopy*. London; San Diego, CA; Cambridge, MA, Oxford: Elsevier.

Medhat, M. E. (2015). A review on applications of the wavelet transform technique in spectral analysis. *J. Appl. Computat. Math.* 4, 1–6. doi: 10.4172/2168-9679.1000224

Neue, G. (1996). Simplification of dynamic NMR spectroscopy by wavelet transform. *Solid State Nucl. Magn. Reson.* 5, 305–314.

Newland, D. E. (1993). *An Introduction to Random Vibrations, Spectral and Wavelet Analysis*. Longman: University of Cambridge. 295–370.

O'Haver, T. C. (1979). Derivative and wavelength modulation spectrometry. *Anal. Chem.* 51, 91A−100A.

O'Haver, T. C., and Green, G. L. (1976). Numerical error analysis of derivative spectrometry for the quantitative analysis of mixtures. *Anal. Chem.* 48, 312–318.

Palavajjhala, S., Motard, R. L., and Joseph, B. (1994). *Wavelet Application in Chemical Engineering*. eds R. L. Motard and B. Joseph (Norwell, MA: Kluwer Academic Publishers), 33–83.

Pektaş, G., Dinç, E., and Baleanu, D. (2009). Combined application of continuous wavelet transform-zero crossing technique in the simultaneous spectrophotometric determination of perindopril and indapamid in tablets. *Quím. Nova* 32, 1416–1421.

Ragno, G., Ioele, G., De Luca, M., Garofalo, A., Grande, F., and Risoli, A. (2006). A critical study on the application of the zero-crossing derivative spectrophotometry to the photodegradation monitoring of lacidipine. *J. Pharm. Biom. Anal.* 42, 39–45. doi: 10.1016/j.jpba.2005.11.025

Salinas, F., Berzas Nevado, J. J., and Espinoso Mansilla, A. E. (1990). A new spectrophotometric method for quantitative multicomponent analysis resolution of mixtures of salicylic and salicyluric acids. *Talanta* 37, 347–351. doi: 10.1016/0039-9140(90)80065-N

Shao, X., Cai, W., and Sun, P. (1998a). Determination of the component number in overlapping multicomponent chromatogram using wavelet transform. *Chemom. Intell. Lab. Syst*. 43, 147–155.

Shao, X., Cai, W., Sun, P., Zhang, M., and Zhao, G. (1997). Quantitative Determination of the components in overlapping chromatographic peaks using wavelet transform. *Anal. Chem.* 69, 1722–1725. doi: 10.1021/ac9608679

Shao, X., Hou, S., Fang, N., He, Y., and Zhao, G. (1998b). Quantitative determination of plant hormones by high performance liquid chromatography with wavelet transform. *Chin. J. Anal. Chem*. 26, 107–110.

Shao, X., Hou, S., and Zhao, G. (1998c). Extraction of the component information from overlapping chromatograms by wavelet transform. *Chin. J. Anal. Chem*. 26, 1428–1431.

Shao, X., and Zhuang, Y. (2004). Determination of chlorogenic acid in plant samples by using near-infrared spectrum with wavelet transform preprocessing. *Anal. Sci*. 20, 451–454. doi: 10.2116/analsci.20.451

Shariati-Rad, M., Irandoust, M., Amini, T., and Ahmadi, F. (2012). Partial least squares and continuous wavelet transformation in simultaneous spectrophotometric determination of amlodipin and atorvastatine. *Pharm. Anal. Acta* 3:178. doi: 10.4172/2153-2435.1000178

Shokry, E., El-Gendy, A. E., Kawy, M. A., and Hegazy, M. (2014). Application of double divisor ratio spectra derivative spectrophotometric [DDRS-DS], chemometric and chromatographic methods for stability indicating determination of moexipril hydrochloride and hydrochlorothiazide. *Curr. Sci. Int*. 3, 352–380.

Sohrabi, M. R., Kamali, N., and Khakpour, M. (2011). Simultaneous spectrophotometric determination of metformin hydrochloride and glibenclamide in binary mixtures using combined discrete and continuous wavelet transforms. *Anal. Sci*. 27, 1037–1041.

Strang, G., and Nguyen, T. (eds.). (1996). Wavelets and Filter Banks. MA: Wellesley-Cambridge Press. 72.

Ugurlu, G., Özaltin, N., and Dinç, E. (2008). Spectrophotometric determination of risedronate sodium in pharmaceutical preparations by derivative and continuous wavelet transforms. *Rev. Anal. Chem.* 27, 215–233. doi: 10.1515/REVAC.2008.27.4.215

Unser, M., and Blu, T. (2000). Fractional splines and wavelets. *SIAM Rev*. 42, 43–67. doi: 10.1137/S0036144598349435

Üstündag, Ö., Dinç, E., and Baleanu, D. (2008). Applications of Mexican hat wavelet function to binary mixture analysis. *Rev. Chim. (Bucuresti)* 59, 1387–1391.

Vetterli, M., and Kovačević, J. (1995). *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice Hall PTR.

Walczak, B., and Massart, D. (1997a). Wavelet packet transform applied to a set of signals: a new approach to the best-basis selection. *Chemom. Intell. Lab. Syst.* 38, 39–50. doi: 10.1016/S0169-7439(97)00050-6

Walczak, B., and Massart, D. L. (1997b). Noise suppression and signal compression using the wavelet packet transform. *Chemom. Intell. Lab. Sys*. 36, 81–94. doi: 10.1016/S0169-7439(96)00077-9

Walczak, B., and Massart, D. L. (1997c). Wavelets-something for analytical chemistry. *Trends Analyt. Chem*. 16, 451–463. doi: 10.1016/S0165-9936(97)00065-4

Walczak, B., and Massart, D. L. (2000a). "Calibration in wavelet domain," in *Wavelets in Chemistry,* ed B. Walczak (Amsterdam: Elsevier), 323–347.

Walczak, B., and Massart, D. L. (2000b). "Joint basis and joint best-basis for data sets," in *Wavelets in Chemistry,* ed B. Walczak (Amsterdam: Elsevier), 165–171.

Walczak, B., and Radomski, J. P. (2000). "Wavelet bases for IR library compression, searching and reconstruction," in *Wavelets in Chemistry,* ed B. Walczak (Amsterdam: Elsevier), 291–308.

Wang, H., Xiao, J., Fan, Z., and Zhang, M. (1997). Wavelet transform and its application in chemistry. *Chem. Bull. Huaxue Tongbao* 6, 20–23.

Wickerhauser, M. V. (1994). *Adapted Wavelet Analysis From Theory to Software*. New York, NY:A.K. Peters, ltd.

Yan-Fang, S. (2013). A review on the applications of wavelet transform in hydrology time series analysis. *Atmospheric Res.*122, 8–15. doi: 10.1016/j.atmosres.2012.11.003

Zheng, X. P., and Mo, J. Y. (1999). The coupled application of the B-spline wavelet and RLT filtration in staircase voltammetry. *Chemometr. Intell. Lab. Syst*. 45, 157-−161. doi: 10.1016/S0169-7439(98)00099-9

Zheng, X. P., Mo, J. Y., and Cai, 'P. X. (1998). Simultaneous application of spline wavelet and Riemann-Liouville transform filtration in electroanalytical chemistry. *Anal. Commun*. 35, 57–59.

Zhong, H. B., Zheng, J. B., Pan, Z. X., Zhang, M. S., and Gao, H. (1998). Investigation on application of wavelet transform in recovering useful information from oscillographic signal. *Chem. J. Chin. Univ*. 19, 547–549.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Chemistry

REVIEW
published: 21 November 2018
doi: 10.3389/fchem.2018.00576

# Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis

*Alessandra Biancolillo and Federico Marini**

*Department of Chemistry, University of Rome La Sapienza, Rome, Italy*

Spectroscopy is widely used to characterize pharmaceutical products or processes, especially due to its desirable characteristics of being rapid, cheap, non-invasive/non-destructive and applicable both off-line and in-/at-/on-line. Spectroscopic techniques produce profiles containing a high amount of information, which can profitably be exploited through the use of multivariate mathematic and statistic (chemometric) techniques. The present paper aims at providing a brief overview of the different chemometric approaches applicable in the context of spectroscopy-based pharmaceutical analysis, discussing both the unsupervised exploration of the collected data and the possibility of building predictive models for both quantitative (calibration) and qualitative (classification) responses.

**Keywords: spectroscopy, chemometrics and statistics, component analysis (PCA), partial least squares (PLS), classification, partial least squares discriminant analysis (PLS-DA), soft independent modeling of class analogies (SIMCA), pharmaceutical quality control**

OPEN ACCESS

***Edited by:***
*Cosimino Malitesta, University of Salento, Italy*

***Reviewed by:***
*Daniel Cozzolino, Central Queensland University, Australia*
*Andreia Michelle Smith-Moritz, University of California, Davis, United States*

***Correspondence:**
*Federico Marini*
*federico.marini@uniroma1.it*

***Specialty section:***
*This article was submitted to Analytical Chemistry, a section of the journal Frontiers in Chemistry*

***Received:*** *07 July 2018*
***Accepted:*** *05 November 2018*
***Published:*** *21 November 2018*

***Citation:***
*Biancolillo A and Marini F (2018) Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. Front. Chem. 6:576. doi: 10.3389/fchem.2018.00576*

## INTRODUCTION

Quality control on pharmaceutical products is undoubtedly an important and widely debated topic. Hence, in literature, various methods have been proposed to check quality of medicines, either qualitative (e.g., for the identification of an active pharmaceutical ingredient, API; Blanco et al., 2000; Herkert et al., 2001; Alvarenga et al., 2008) or quantitative (quantification of the API; Blanco et al., 2000; Yao et al., 2007; Cruz Sarraguça and Almeida Lopes, 2009); involving either destructive or non-invasive online techniques. Recently, due to the benefits they bring, several non-destructive methodologies based on spectroscopic techniques (mainly Near-Infrared NIR) combined with chemometric tools have been proposed for pharmaceutical quality check (Chen et al., 2018; Rodionova et al., 2018).

Despite the development of analytical methodologies and the commitments of national and supranational entities to regulate pharmaceutical quality control, substandard and counterfeit medicines are still a major problem all over the world.

### Chemometrics as Tool for Fraud/Adulteration Detection

Poor-quality pharmaceuticals can be found on the market for two main reasons: low production standards (mainly leading to substandard medicines) and fraud attempts. Counterfeited drugs may present different frauds/adulterations; for instance, they could contain no active pharmaceutical ingredient (API), a different API from the one declared, or a different (lower) API strength. As mentioned above, several methodologies have been proposed in order to

detect substandard/counterfeit pharmaceuticals; among these, a major role is played by those based on the application of spectroscopic techniques in combination with different chemometric methods. The relevance of these methodologies is due to the fact that spectroscopy (in particular, NIR) combined with exploratory data analysis, classification and regression method can lead to effective, high performing, fast, non-destructive, and sometimes, online methods for checking the quality of pharmaceuticals and their compliance to production and/or pharmacopeia standards. Nevertheless, the available chemometric tools applicable to handle spectroscopic (but, of course not only those) data are numerous, and there is plenty of room for their misapplication (Kjeldahl and Bro, 2010). As a consequence, the aim of the present paper is to report and critically discuss some of the chemometric methods typically applied for pharmaceutical analysis, together with an essential description of the figures of merit which allow evaluating the quality of the corresponding models.

## EXPLORATORY DATA ANALYSIS

In the large part of the studies for the characterization of pharmaceutical samples for quality control, verification of compliance and identification/detection of counterfeit, fraud or adulterations, experimental signals (usually in the form of some sorts of fingerprints) are collected on a series of specimens. These constitute the data the chemometric models operate on. These data are usually arranged in the form of a matrix $\boldsymbol{X}$, having as many rows as the number of samples and as many columns as the number of measured variables. Accordingly, assuming that samples are spectroscopically characterized by collecting an absorption (or reflection/transmission) profile (e.g., in the infrared region), each row of the matrix corresponds to the whole spectrum of a particular sample, whereas each column represents the absorbance (or reflectance/transmittance) of all the individuals at a particular wavenumber. This equivalence between the experimental profiles and their matrix representation is graphically reported in **Figure 1**.

Once the data have been collected, exploratory data analysis represents the first step of any chemometric processing, as it allows "to summarize the main characteristics of data in an easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis" (Tukey, 1977). Exploratory data analysis provides an overall view of the system under study, allowing to catch possible similarities/dissimilarities among samples, to identify the presence of clusters or, in general, systematic trends, to discover which variables are relevant to describe the system and, on the other hand, which could be in principle discarded, and to detect possible outlying, anomalous or, at least, suspicious samples (if present). As evident also from the definition reported above, in the context of exploratory data analysis a key role is played by the possibility of capturing the main structure of the data in a series of representative plots, through appropriate display techniques. Indeed, considering a general data matrix $\boldsymbol{X}$, of dimensions $N \times M$, one could think of its entries as the coordinates of $N$ points (the samples) into a $M$-dimensional space whose axes are the variables, which makes this representation unfeasible for the cases when more than three descriptors are collected on each individual. This is why exploratory data analysis often relies on the use of projection (bilinear) techniques to reduce the data dimensionality in a "clever" way. Projection methods look for a low-dimensional representation of the data, whose axes (normally deemed components or latent variables) are as relevant as possible for the specific task. In the case of exploratory data analysis, the most commonly used technique is Principal Components Analysis (PCA) (Pearson, 1901; Wold et al., 1987; Jolliffe, 2002).

### Principal Component Analysis

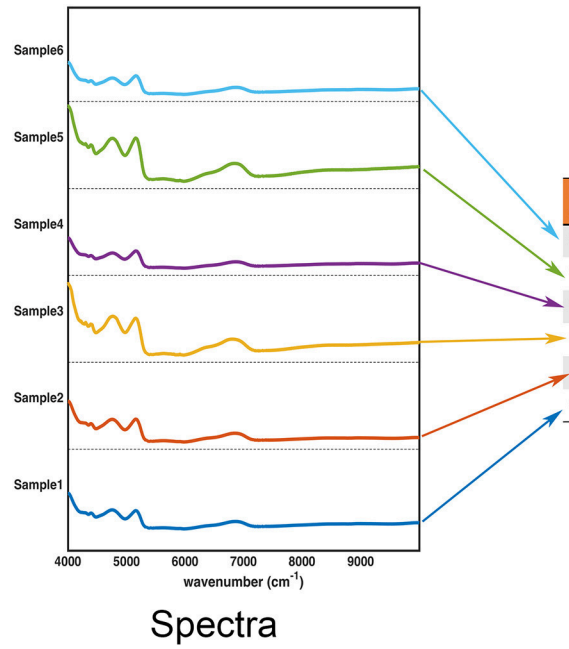
Principal component analysis (PCA) is a projection method, which looks for directions in the multivariate space progressively providing the best fit of the data distribution, i.e., which best approximate the data in a least squares sense. This explains why PCA is the technique of choice in the majority of cases when exploratory data analysis is the task: indeed, by definition, for any desired number of dimensions (components) $F$ in the final representation, the subspace identified by PCA constitutes the most faithful $F$-dimensional approximation of the original data. This allows compression of the data dimensionality at the same time reducing to a minimum the loss of information. In particular, starting from a data matrix $\boldsymbol{X}_{(N \times M)}$, Principal Component Analysis is based on its bilinear decomposition, which can be mathematically described by Equation (1):

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{1}$$

The *loadings* matrix $\boldsymbol{P}_{(M \times F)}$ identifies the $F$ directions, i.e., the *principal components* (PC), along which the data should be projected and the results of such projection, i.e., the coordinates of the samples onto this reduced subspace, are collected in the *scores matrix* $\boldsymbol{T}_{(N \times F)}$. In order to achieve data compression, usually $F \ll M$ so that the PCA representation provides an approximation of the original data whose residuals are collected in the matrix $\boldsymbol{E}_{(N \times M)}$.

Since the scores represent a new set of coordinates along highly informative (relevant) directions, they may be used in two- or three-dimensional scatterplots (scores plots). This offers a straightforward visualization of the data, which can highlight possible trends in data, presence of clusters or, in general, of an underlying structure. A schematic representation of how PCA works is displayed in **Figure 2**.

**Figure 2** shows one of the simplest possible examples of feature reduction, since it describes the case where samples described by three measured variables can be approximated by being projected on an appropriately chosen two-dimensional sub-space. However, the concept may be easily generalized to higher-dimensional problems, such as those involving spectroscopic measurements. **Figure 3** shows an example of the application of PCA to mid infrared spectroscopic data. In particular, the possibility of extracting as much information as possible from the IR spectra recorded on 51 tablets containing




data matrix **X**

| log(1/R)@ 4000 cm$^{-1}$ | log(1/R)@ 4004 cm$^{-1}$ | log(1/R)@ 4008 cm$^{-1}$ | log(1/R)@ 4012 cm$^{-1}$ | .... | log(1/R)@ 9992 cm$^{-1}$ | log(1/R)@ 9996 cm$^{-1}$ | log(1/R)@ 10000 cm$^{-1}$ |
|---|---|---|---|---|---|---|---|
| 0.4893 | 0.4894 | 0.4891 | 0.4895 | .... | 0.0836 | 0.0830 | 0.0828 |
| 0.5618 | 0.5617 | 0.5617 | 0.5618 | .... | 0.0601 | 0.0604 | 0.0606 |
| 0.9817 | 0.9847 | 0.9871 | 0.9880 | .... | 0.1725 | 0.1725 | 0.1731 |
| 0.4172 | 0.4171 | 0.4166 | 0.4164 | .... | 0.0701 | 0.0699 | 0.0694 |
| 1.0375 | 1.0390 | 1.0390 | 1.0385 | .... | 0.2018 | 0.2016 | 0.2026 |
| 0.4397 | 0.4395 | 0.4392 | 0.4396 | .... | 0.0758 | 0.0757 | 0.0753 |

**FIGURE 1 |** Graphical illustration of the equivalence between the collected experimental data (in this case, NIR spectra for 6 samples) and the data matrix **X**. Each row of the data matrix corresponds to the spectrum of a sample, whereas each column contains the value of a specific variable over all the individuals.

either ketoprofen or ibuprofen in the region 2,000–680 cm$^{-1}$ (661 variables) is represented.

A large portion of the data variability can be summarized by projecting the samples onto the space spanned by the first two principal components, which account for about 90% of the original variance, and therefore can be considered as a good approximation of the experimental matrix. Inspection of the scores plot suggests that the main source of variability is the difference between ibuprofen tablets (blue squares) and ketoprofen ones (red circles), since the two clusters are completely separated along the first principal component. To interpret the observed cluster structure in terms of the measured variables, it is then necessary to inspect the corresponding loadings, which are also displayed in **Figure 3** for PC1. Indeed, for spectral data, the possibility of plotting the loadings for the individual components in a profile-like fashion, rather than producing scatterplot for pairs of latent variables (as exemplified in **Figure 2**) is often preferred, due to its more straightforward interpretability: spectral regions having positive loadings will have higher intensity on samples which have positive scores on the corresponding component, whereas bands associated to negative loadings will present higher intensity on the individuals falling at negative values of the PC. In the example reported in **Figure 3**, one could infer, for instance, that the ketoprofen samples (which fall at positive values of PC1) have a higher absorbance at the wavenumbers where the loadings are positive, whereas ibuprofen samples should present a higher signal in correspondence to the bands showing negative loadings.

Based on what reported above, it is evident how the quality of the compressed representation in the PC space depends on the number of components $F$ chosen to describe the data. However, at the same time, it must be noted that when the aim of calculating a PCA is "only" data display, as in most of the applications in the context of exploratory analysis, the choice of the optimal number of components is not critical: it is normally enough to inspect the data distribution across the first few dimensions and, in many cases, considering the scores plot resulting from the first two or three components could be sufficient. On the other hand, there may be cases when the aim of the exploratory analysis is not limited to just data visualization and, for instance, one is interested in the identification of anomalous or outlying observations, or there could be the need of the imputation of missing elements in the data matrix; additionally, one could also need to obtain a compressed representation of the data to be used for further predictive modeling. In all such cases, the choice of the optimal dimensionality of the PC representation is critical for the specific purposes and, therefore, the number of PCs should be carefully estimated. In this respect, different methods have been proposed in the literature and a survey of the most commonly used can be found in Jolliffe (2002).

Among the applications described above, the possibility of using PCA for the identification/detection of potential outliers deserves a few more words, as it could be of interest for pharmaceutical quality control. Actually, although outliers—or anomalous observations, in general—could be, in principle, investigated by visually inspecting the scores plot along the first components, this approach could be subjective and anyway would not consider some possible data discrepancies. Alternatively, when it is used as a model to

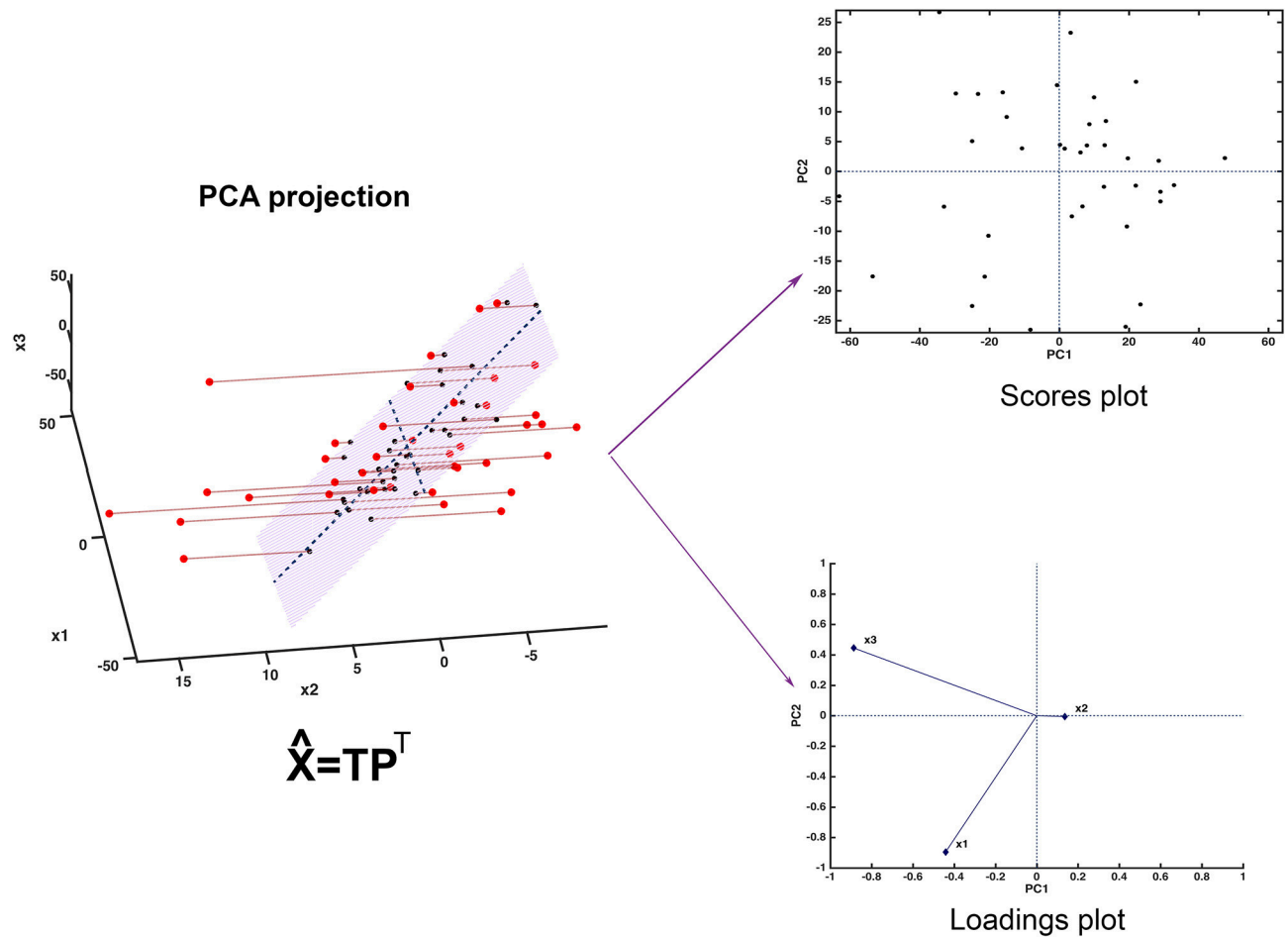



**FIGURE 2 |** Graphical illustration of the basics of PCA. The samples, here represented in a three-dimensional space, are projected onto a low-dimensional subspace (highlighted in light red in the leftmost panel) spanned by the first two principal components. Inspection of the data set can be carried out by looking at the distribution of the samples onto the informative PC subspace (scores plot) and interpretation can be then carried out by examining the relative contribution of the experimental variable to the definition of the principal components (loadings plot).

build a suitable approximation of the data, PCA provides a powerful toolbox for outlier detection based on the definition of more objective test statistics, which can be easily automatized or, anyway, embedded in control strategies, also on-line. This is accomplished by defining two distance measurements: (i) a squared Mahalanobis distance in the scores space, which follows the $T^2$ statistics (Hotelling, 1931) and accounts for how extreme the measurement is in the principal component subspace, and (ii) a squared orthogonal Euclidean distance (the sum of squares of the residuals after approximating the observation by its projection), which is normally indicated as Q statistics (Jackson and Muldholkar, 1979) and quantifies how well the model fits that particular individual. Outlier detection is then carried out by setting appropriate threshold values for the $T^2$ and Q statistics and verifying whether the samples fall below or above those critical limits. Moreover, once an observation is identified as a potential outlier, inspection of the contribution plot can help in relating the detected anomaly to the behavior of specific measured variables.

## Selected Examples

PCA is customarily used for the quality control of drugs and pharmaceuticals; several examples of the application of this technique to solve diverse issues have been reported in the literature. One of the most obviously relevant ones is fraud detection. For example, in Rodionova et al. (2005) PCA was applied to both bulk NIR spectroscopy and hyperspectral imaging (HSI) in the NIR range to spot counterfeit drugs. In particular, bulk NIR was used to differentiate genuine antispasmodic drugs from forgeries, whereas HSI on the ground uncoated tablets was employed to identify counterfeited antimicrobial drugs. In both cases, the spectroscopic data were subjected to PCA, which allowed to clearly identify clusters in the scores plot, corresponding to the two kinds of tablets, i.e., genuine and counterfeited. In the case of the imaging platform, where the signal is stored as a data hypercube [i.e., a three-way numerical array of dimension number of horizontal pixels $N_x$, number of vertical pixels $N_y$ and number of wavelengths $N_\lambda$, in which each entry corresponds to the spectral intensity measured

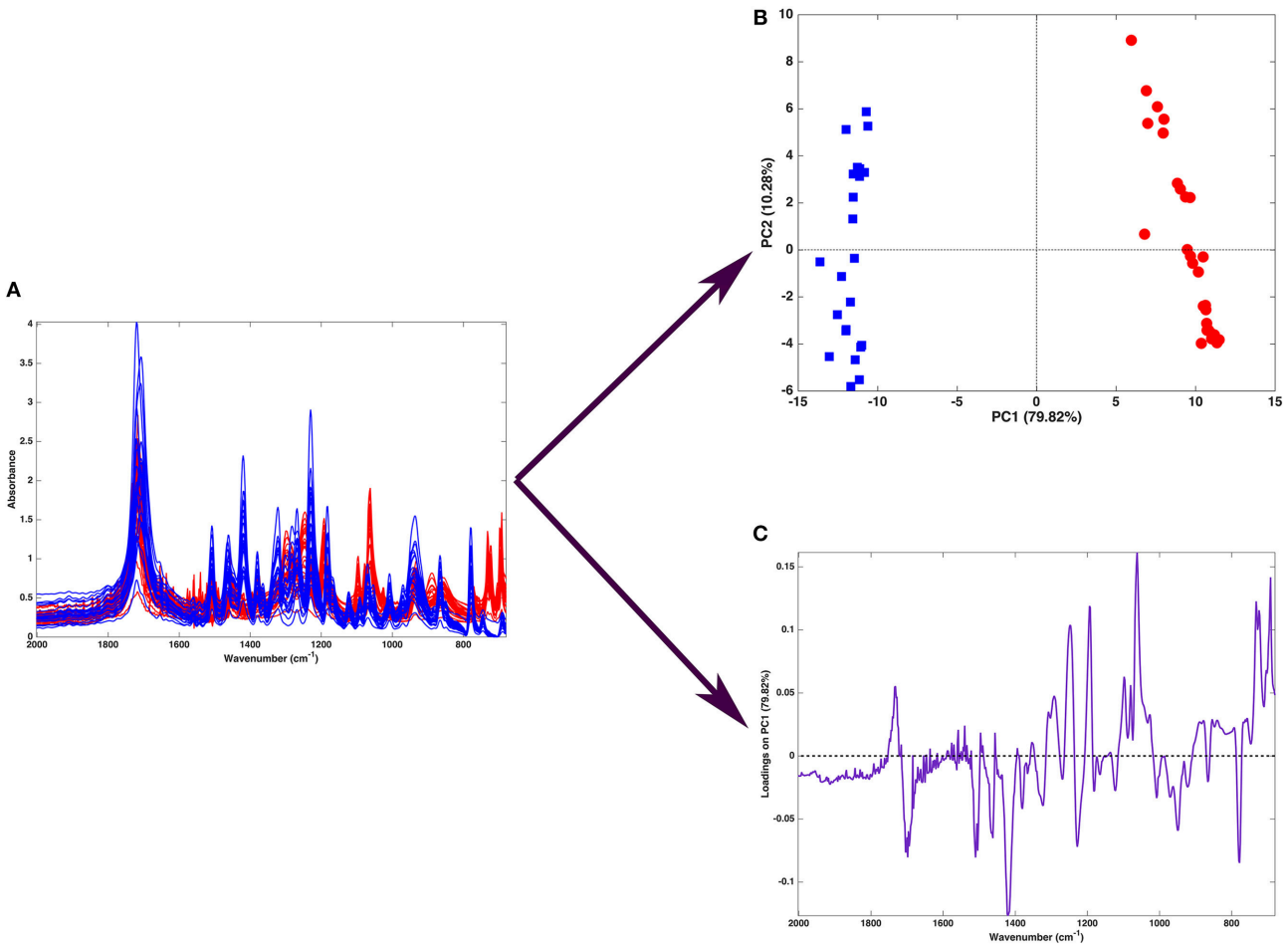



**FIGURE 3 |** Graphical illustration of the application of PCA on a spectral (mid-infrared) data. Fifty-one spectra recorded on samples containing either ibuprofen (blue) or ketoprofen (red) are recorded in the region 680–2,000 $cm^{-1}$ **(A)**. When PCA is applied to such a dataset, one obtains a scores plot **(B)** showing that two cluster of samples, corresponding to tablets containing ibuprofen (blue squares) or ketoprofen (red circles) are separated along the first component. Interpretation of the observed differences in terms of the spectroscopic signal is made possible by the inspection of the loadings on PC, which are shown in a "spectral-like" fashion in **(C)**.

at a certain wavelength and a specific spatial position (x-y coordinates)], a preliminary unfolding step is needed. Unfolding is the procedure allowing to reorganize a higher-order array into a two-way matrix, which can be then processed with standard chemometric techniques. In the case of hyperspectral data cubes, this is carried out by stacking the spectra corresponding to the different pixels one on top of each other, in a way to obtain a matrix of dimensions ($N_x \times N_y$ and $N_\lambda$).

Another relevant application of exploratory analysis is related to quality check. For instance, PCA can be applied to investigate formulations not meeting predefined parameters. In Roggo et al. (2005), PCA was used to inquire a suspicious blue spot present on tablets. Samples were analyzed by a multi-spectral (IR) imaging microscope and PCA analysis was performed on the unfolded data-cube, indicating that the localized coloration was not due to contamination, but actually given by wet indigo carmine dye and placebo (expected ingredients of the formulation).

PCA can also be used for routine quality checks at the end of a production process. For example, in Myakalwar et al. (2011) laser-induced breakdown spectroscopy (LIBS) and PCA were combined with the aim of obtaining qualitative information about the composition of different pharmaceuticals.

## REGRESSION

As discussed in the previous section, exploratory analysis is a first and fundamental step in chemometric data processing and, in some cases, it could be the only approach needed to characterize the samples under investigation. However, due to its unsupervised nature, it provides only a (hopefully) unbiased picture of the data distribution but it lacks any possibility of formulating predictions on new observations, which on the other hand may be a fundamental aspect to solve specific issues. In practice, very often quality control and/or authentication of pharmaceutical products rely on some forms of qualitative or quantitative predictions. For instance, the quantification of a specific compound (e.g., an active ingredient or an

excipient) contained in a formulation is a routine operation in pharmaceutical laboratories. This goal can be achieved by combining instrumental (e.g. spectroscopic) measurements with chemometric regression approaches (Martens and Naes, 1991; Martens and Geladi, 2004). Indeed, given a response to be predicted $y$ and a vector of measured signals (e.g., a spectrum) $\boldsymbol{x}$, the aim of regression methods is to find the functional relationship that best approximates the response on the basis of the measurements (the *predictors*). Mathematically, this can be stated as:

$$y = \hat{y} + e = f(\boldsymbol{x}) + e \tag{2}$$

where $\hat{y}$ is the predicted response (i.e., the response value approximated by the model), $f(\boldsymbol{x})$ indicates a general function of and $\boldsymbol{x}$ and $e$ is the residual, i.e., the difference between the actual response and its predicted value. In many applications, the functional relationship between the response and the predictors $f(\boldsymbol{x})$ can be assumed to be linear:

$$\hat{y} = f(\boldsymbol{x}) = b_1 x_1 + b_2 x_2 + \ldots + b_M x_M = \boldsymbol{x}^T \boldsymbol{b} \tag{3}$$

where $x_1, x_2 \ldots x_M$ are the components of the vector of measurements $\boldsymbol{x}$ and the transpose indicates that it is normally expressed as a row vector, while the associated linear coefficients $b_1, b_2 \ldots b_M$, which weight the contributions of each of the $M$ $X$-variables to $y$, are called regression coefficients and collected in the vector $\boldsymbol{b}$. Building a regression model means to find the optimal value of the parameters $\boldsymbol{b}$, i.e., the values which lead to the lowest error in the prediction of the responses. As a direct consequence of this consideration, it is obvious how it is mandatory to have a set of samples (the so-called *training set*) for which both the experimental data $\boldsymbol{X}$ and the responses $\boldsymbol{y}$ are available, in order to build a predictive model. Indeed, the information on the $\boldsymbol{y}$ is actively used to calculate the model parameters. When data from more than a single sample are available, the regression problem in Equations (2, 3) can be reformulated as:

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e} \tag{4}$$

where the vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ collect the predictions and residuals for the different samples, respectively. Accordingly, the most straightforward way of calculating the model parameters in Equation (4) is by the ordinary least-squares approach, i.e., by looking at those values of $\boldsymbol{b}$, which minimize the sum of squares of the residuals $\boldsymbol{e}$:

$$\min_{\boldsymbol{b}} \boldsymbol{e}^T \boldsymbol{e} = \min_{\boldsymbol{b}} \sum_{i=1}^{N} e_i^2 \tag{5}$$

$e_i$ being the residual for the i$^{\text{th}}$ sample and $N$ being the number of training observations. The corresponding methods is called multiple linear regression (MLR) and, under the conditions of Equation (5), the regression coefficients are calculated as:

$$\boldsymbol{b} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{6}$$

Here it is worth to highlight that, if one wishes to use the same experimental matrix $\boldsymbol{X}$ to predict more than one response, i.e., if, for each sample, instead of a single scalar $y_i$, there is a dependent vector

$$\boldsymbol{y}_i^T = \left[ y_{i1} y_{i2} \cdots y_{iL} \right] \tag{7}$$

$L$ being the number of responses, then each dependent variable should be regressed on the independent block by means of a set of regression coefficients. Assuming that the $L$ responses measured on the training samples are collected in a matrix $\boldsymbol{Y}$, whose columns $\boldsymbol{y}_l$ are the individual dependent variables,

$$\boldsymbol{Y} = \left[ \boldsymbol{y}_1 \cdots \boldsymbol{y}_l \cdots \boldsymbol{y}_L \right] \tag{8}$$

the corresponding regression equations could be written as:

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}\boldsymbol{b}_1 + \boldsymbol{e}_1 \\ &\vdots \\ \boldsymbol{y}_l &= \boldsymbol{X}\boldsymbol{b}_l + \boldsymbol{e}_l \\ &\vdots \\ \boldsymbol{y}_L &= \boldsymbol{X}\boldsymbol{b}_L + \boldsymbol{e}_L \end{aligned} \tag{9}$$

which can be grouped into a single expression:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{10}$$

where the residuals, i.e., the differences between the measured and predicted responses are collected in the matrix $\boldsymbol{E}$, and the regression coefficients vectors are gathered in a matrix $\boldsymbol{B}$, which can be estimated, analogously to Equation (6), as:

$$\boldsymbol{B} = [\boldsymbol{b}_1 \cdots \boldsymbol{b}_l \cdots \boldsymbol{b}_L] = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{11}$$

Equations (9–11) indicate that, as far as MLR is concerned, building a model to predict one response at a time or another model to predict multiple responses altogether would lead to the same results since, in the latter case, each dependent variable is anyway modeled as if it were alone. In either case, the solutions of the least-squares problem reported in Equations (6, 11) rely on the possibility of inverting the matrix $\left(\boldsymbol{X}^T \boldsymbol{X}\right)$, i.e., on the characteristics of the predictors. Indeed, in order for that matrix to be invertible, the number of samples should be higher than that of variables and the variables themselves should be as uncorrelated as possible. These conditions are rarely met by the techniques which are used to characterize pharmaceutical samples and, in particular, never met by spectroscopic methods. Due to these limitations, alternative approaches have been proposed in the literature to build regression models in cases where standard multiple linear regression is not applicable. In particular, since in order for the regression solution to exist, the predictor matrix should be made of few, uncorrelated variables, most of the alternative approaches proposed in the literature involve the projection of the $\boldsymbol{X}$ matrix onto a reduced space of orthogonal components and the use of the corresponding scores as regressors to predict the response(s). In this regard, one of the most widely used approaches is principal component regression (PCR) (Hotelling, 1957; Kendall, 1957; Massy, 1965; Jeffers, 1967; Jolliffe, 1982, 2002; Martens and Naes, 1991;

Martens and Geladi, 2004) which, as the name suggests, involves a two-stage process where at first principal component analysis is used to compress the information in the $\boldsymbol{X}$ block onto a reduced set of relevant scores, as already described in Equation (1):

$$\boldsymbol{T} = \boldsymbol{XP} \tag{12}$$

and then these scores constitute the predictor block to build a multiple linear regression:

$$\hat{\boldsymbol{Y}} = \boldsymbol{TC} \tag{13}$$

$\boldsymbol{C}$ being the matrix of regression coefficients for this model. By combining Equations (12, 13), it can be easily seen how PCR still describes a linear relationship between the responses $\boldsymbol{Y}$ and the original variables $\boldsymbol{X}$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{TC} = \boldsymbol{XPC} = \boldsymbol{XB}_{PCR} \tag{14}$$

mediated by a matrix of regression coefficients $\boldsymbol{B}_{PCR}$ ($=\boldsymbol{PC}$), which is different from the one that would be estimated by Equation (11), since it is calculated by taking into account only the portion of the variability in the $\boldsymbol{X}$ block accounted for by the selected principal components. The use of principal component scores as predictors allows to solve the issues connected to the matrix $\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ being usually ill-conditioned when dealing with spectroscopic techniques, but may be still suboptimal in terms of predictive accuracy.

Indeed, as described in Equations (12, 13), calculating a PCR model is a two-step procedure, which involves at first the calculation of PC scores and then the use of these scores to build a regression model to predict the response(s). However, these two steps have different objective functions, i.e., the criterion which is used to extract the scores from the $\boldsymbol{X}$ matrix is not the same which guides the calculation of the regression coefficients $\boldsymbol{C}$ in Equation (13). Stated in different words, the directions of maximum explained variance (especially when there are many uninformative sources of variability in the data) may not be relevant for the prediction of the $\boldsymbol{Y}$. To overcome this drawback, an alternative approach to component-based regression is represented by the Partial Least-Squares algorithm (Wold et al., 1983; Geladi and Kowalski, 1986; Martens and Naes, 1991) which, due to its being probably the most widely used calibration method in chemometrics, will be described in greater detail in the following subparagraph.

## Partial Least Squares (PLS) Regression

Partial Least Squares (PLS) regression (Wold et al., 1983; Geladi and Kowalski, 1986; Martens and Naes, 1991) was proposed as an alternative method to calculate reliable regression models in the presence of ill-conditioned matrices. Analogously to PCR, it is based on the extraction of a set of scores $\boldsymbol{T}$ by projecting the $\boldsymbol{X}$ block on a subspace of latent variables, which are relevant for the calibration problem. However, unlike PCR, the need for the components not only to explain a significant portion of the $\boldsymbol{X}$ variance but also to be predictive for the response $\boldsymbol{Y}$ is explicitly taken into account for the definition of the scores. Indeed, in PLS, the latent variables (i.e., the directions onto which the data are projected) are defined in such a way to maximize the covariance between the corresponding scores and the response(s): maximizing the covariance allows to obtain scores which at the same time describe a relevant portion of the $\boldsymbol{X}$ variance and are correlated with the response(s). Due to these characteristics, and differently than what already described in the case of MLR (see Equation 11) and, by extension, PCR, in PLS two distinct algorithms have been proposed depending on whether only one or multiple responses should be predicted (the corresponding approaches are named PLS1 and PLS2, respectively). In the remainder of this section, both algorithms will be briefly described and commented.

When a single response has to be predicted, its values on the training samples are collected in a vector $\boldsymbol{y}$; accordingly, the PLS1 algorithm extracts scores from the $\boldsymbol{X}$ block having maximum covariance with the response. In particular, the first score $\boldsymbol{t}_1$ is the projection of the data matrix $\boldsymbol{X}$ along the direction of maximum covariance $\boldsymbol{r}_1$:

$$\max_{\boldsymbol{r}_1}\left[Cov(\boldsymbol{t}_1, \boldsymbol{y})\right] = \max_{\boldsymbol{r}_1}\left(\boldsymbol{t}_1^T\boldsymbol{y}\right) \tag{15}$$

While the successive scores $\boldsymbol{t}_2 \cdots \boldsymbol{t}_F$, which are all orthogonal, account in turn for the maximum residual covariance. Therefore, PLS1 calculates a set of orthogonal scores having maximum covariance with $\boldsymbol{y}$, according to:

$$\boldsymbol{T} = \boldsymbol{XR} \tag{16}$$

$\boldsymbol{R}$ being the weights defining the subspace onto which the matrix should be projected, and then uses these scores as regressors for the response:

$$\hat{\boldsymbol{y}} = \boldsymbol{Tq} \tag{17}$$

$\boldsymbol{q}$ being the coefficients for the regression. Similarly to what already shown in the case of PCR, Equations (16, 17) can be then combined in a single one to express the regression model as a function of the original variables, through the introduction of the regression vector $\boldsymbol{b}_{PLS1}$ ($=\boldsymbol{Rq}$):

$$\hat{\boldsymbol{y}} = \boldsymbol{Tq} = \boldsymbol{XRq} = \boldsymbol{Xb}_{PLS1}. \tag{18}$$

In contrast, in the multi-response case (PLS2), it is assumed that also the matrix $\boldsymbol{Y}$, which collects the values of the dependent variables on the training samples, has a latent structure, i.e., it can be approximated by a component model:

$$\hat{\boldsymbol{Y}} = \boldsymbol{UQ}^T \tag{19}$$

$\boldsymbol{U}$ and $\boldsymbol{Q}$ being the $\boldsymbol{Y}$ scores and loadings, respectively. In particular, in order for the calibration model to be efficient, it is assumed that the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ matrices share the same latent structure. This is accomplished by imposing that the component be relevant to describe the variance of the independent block and predictive for the responses. In mathematical terms, pairs of

scores are simultaneously extracted from the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ blocks so to have maximum covariance:

$$\max_{\boldsymbol{r}_i,\boldsymbol{q}_i}[Cov(\boldsymbol{t}_i,\boldsymbol{u}_i)] = \max_{\boldsymbol{r}_i,\boldsymbol{q}_i}(\boldsymbol{t}_i^T\boldsymbol{u}_i) \quad (20)$$

Where $\boldsymbol{t}_i$ and $\boldsymbol{u}_i$ are the $\boldsymbol{X}$ and the $\boldsymbol{Y}$ scores on the ith latent variable, respectively, $\boldsymbol{q}_i$ is the ith column of the $\boldsymbol{Y}$ loading matrix $\boldsymbol{Q}$ while $\boldsymbol{r}_i$ is the ith column of the $\boldsymbol{X}$ weight matrix $\boldsymbol{R}$, which has the same meaning as specified in Equation (16). Additionally, these scores are made to be collinear, through what is normally defined as the inner relation:

$$\boldsymbol{u}_i = \boldsymbol{t}_i c_i \ \forall i \quad (21)$$

$c_i$ being a proportionality constant (*inner regression coefficient*). When considering all the pairs of components, Equation (21) can be rewritten in a matrix form as:

$$\boldsymbol{U} = \boldsymbol{TC} \quad (22)$$

where:

$$\boldsymbol{C} = \begin{bmatrix} c_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & c_F \end{bmatrix}. \quad (23)$$

Also in this case, by combining all the equations defining the model, it is possible to express the predicted responses as a linear function of the original variables:

$$\hat{\boldsymbol{Y}} = \boldsymbol{UQ}^T = \boldsymbol{TCQ}^T = \boldsymbol{XRCQ}^T = \boldsymbol{XB}_{PLS2} \quad (24)$$

where the matrix of regression coefficients $\boldsymbol{B}_{PLS2}$ is defined as $\boldsymbol{RCQ}^T$.

Based on the above description, it is clear that, when more than one response has to be modeled, it is essential to decide whether it could be better to build an individual model for each dependent variable, or a single model to predict all the responses, as the results would not be the same. In particular, it is advisable to use the PLS2 approach only when one could reasonably assume that there are systematic relationships between the dependent variables.

On the other hand, independently on what model one decides to use, once the values of the regression coefficients (here generally indicated as $\boldsymbol{B}$) have been estimated based on the training samples, they can be used to predict the responses for any new set of measurements ($\boldsymbol{X}_{new}$):

$$\hat{\boldsymbol{Y}}_{new} = \boldsymbol{X}_{new}\boldsymbol{B}. \quad (25)$$

Here, it should be stressed that, in order for the calibrations built by PLS (but the same concept holds for PCR) to be accurate and reliable, a key parameter is the choice of an appropriate number of latent variables to describe the data. Indeed, while selecting a low number of components one can incur in the risk of not explaining all the relevant variance (*underfitting*), including too many of them (so that not only the systematic information is captured, but also the noise), can lead to *overfitting*, i.e., to a model which is very good in predicting the samples it has been calculated on, but performs poorly on new observations. To reduce this risk, a proper validation strategy is needed (see section Validation) and, in particular, the optimal number of latent variables is selected as the one leading to the minimum error during one of the validation stages (usually, cross-validation).

## Selected Application of Regression Methods to Pharmaceutical Problems

Regression methods in general, and especially PLS, are often combined with spectroscopy in order to develop rapid and (sometimes) non-destructive methodologies for the quantification of active ingredients in formulations. For instance, Bautista et al. (1996) quantified three analytes of interest (caffeine, acetylsalicylic acid and acetaminophen) in their synthetic ternary mixtures and different formulations by UV-Vis spectroscopy assisted by a PLS calibration model. Mazurek et al. proposed two approaches based on coupling FT-Raman spectroscopy with PLS and PCR calibration for estimation of captopril and prednisolone in tablets (Mazurek and Szostak, 2006a) and diclofenac sodium and aminophylline in injection solutions (Mazurek and Szostak, 2006b). The authors compared results obtained from calibration models built by using unnormalised spectra with the values found when an internal standard was added to each sample and the spectra were normalized by its selected band intensity at maximum or integrated. Another study on injection samples was proposed by Xie et al. (2010), using NIR spectroscopy combined with PLS and PCR to quantify pefloxacin mesylate (an antibacterial agent) in liquid formulations. PLS regression was also coupled to MIR (Marini et al., 2009) and NIR spectroscopy (Rigoni et al., 2014) to quantify the enantiometric excess of different APIs in the solid phase, also in the presence of excipients, based on the consideration that, in the solid phase, the spectrum of the racemic mixture could be different from that of either pure enantiomer. Specifically, it was possible to accurately quantify the enantiomeric excess of S-(+)-mandelic acid and S-(+)-ketoprofen by MIR spectroscopy coupled by PLS on the whole spectrum and after variable selection by sequential application of backward interval PLS and genetic algorithms (biPLS-GA) (Marini et al., 2009), while NIR was used to quantify the enantiomeric excess of R-(−)-epinephrine and S-(+)-ibuprofen (Rigoni et al., 2014). In the latter case, it was also shown that, when using the validated model to quantify the enantiomeric excess of API in the finished products, the influence of excipients and dosage forms (intact tablets or powders) has a relevant impact on the final predictive accuracy.

# CLASSIFICATION

As already introduced in the previous section, in chemometric applications, in general, and in the context of pharmaceutical analysis, in particular, one is often interested in using the experimentally collected data (e.g., spectroscopic profiles) to predict qualitative or quantitative properties of the samples.

While the regression methods for the prediction of quantitative responses have been already presented and discussed in section Regression, the main chemometric approaches for the prediction of qualitative properties of the individuals under investigation are outlined herein. These approaches are generally referred to as classification methods, since any discrete level that the qualitative variable can assume may also be defined as a class (or category) (Bevilacqua et al., 2013). For instance, if one were interested in the possibility of recognizing which of three specific sites a raw material was supplied from, it is clear that the response to be predicted could only take three possible values, namely "Site A," "Site B," and "Site C"; each of these three values would correspond to a particular class. A class can be then considered as an ensemble of individuals (samples) sharing similar characteristics. In this example, samples from the first class would all be characterized by having been manufactured from a raw material produced in Site A, and similar considerations could be made for the specimens in the second and third classes, corresponding to Site B and Site C, respectively. As it could already be clear from the example, there are many ambits of application for classification methods in pharmaceutical and biomedical analysis, some of which will be further illustrated in section Selected Applications of Classification Approaches for Pharmaceutical Analysis, after a brief theoretical introduction to the topic as well as the chemometric methods most frequently used in this context (especially, in combination with spectroscopic techniques).

As anticipated above, classification approaches aim at relating the experimental data collected on a sample to a discrete value of a property one wishes to predict. This same problem can be also expressed in geometrical terms by considering that each experimental profile (e.g., spectrum) can be seen as point in the multivariate space described by the measured variables. Accordingly, a classification problem can be formulated as the identification of regions in this multivariate space, which can be associated to a particular category, so that if a point falls in one of these regions, it is predicted as being part of the corresponding class. In this respect, classification approaches can be divided into two main sub-groups: discriminant and class-modeling methods. In this framework, a fundamental distinction can be made between discriminant and class-modeling tools, which constitute the two main approaches to perform classification in chemometrics (Albano et al., 1978). In detail, discriminant classification methods focus on identifying boundaries in the multivariate space, which separate the region(s) corresponding to a particular category from those corresponding to another one. This means they need representative samples from all the categories of interest in order to build the classification model, which will be then able to predict any new sample as belonging only to one of the classes spanned by the training set. In a problem involving three classes, a discriminant classification method will look for those boundaries in the multivariate space identifying the regions associated to the three categories in such a way as to minimize the classification error (i.e., the percentage of samples wrongly assigned). An example is reported in **Figure 4A**. On the other hand, class-modeling techniques look at the similarities among individuals belonging to the same category, and aim at defining a (usually bound) subspace where samples from the class under investigation can be found with a certain probability; in this sense, they resemble outlier tests, and indeed they borrow most of the machinery from the latter. Operationally, each category is modeled independently on the others and the outcome is the definition of a class boundary which should enclose the category sub-space:, i.e., individuals falling within that space are likely to belong to the class (are "accepted" by the class model), whereas samples falling outside are deemed as outliers and rejected. It is then evident that one of the main advantages of class modeling approaches is that they allow building a classification model also in the asymmetric case, where there is only a category of interest and the alternative one is represented by all the other individuals not falling under the definition of that particular class. In this case, since the alternative category is ill-defined, heterogeneous, and very likely to be underrepresented in the training set, any discriminant model would result suboptimal, as its predictions would strongly depend on the (usually not enough) samples available for that class. On the other hand, modeling techniques define the category space only on the basis of data collected for the class of interest, so those problems can be overcome.

When the specific problem requires to investigate more than one class, each category is modeled independently on the others and, accordingly, the corresponding sub-spaces may overlap (see **Figure 4B**). As a consequence, classification outcomes are more versatile than with discriminant methods: a sample can be accepted by a single category model (and therefore be assigned to that class), by more than one (falling in the area where different class spaces overlap and, hence, resulting "confused") or it could fall outside any class-region and therefore be rejected by all the categories involved in the model.

## Discriminant Methods

As mentioned above, predictions made by the application of discriminant methods are univocal; namely, each sample is uniquely assigned to one and only one of the classes represented in the training set. This is accomplished by defining decision surfaces, which delimit the boundaries among the regions of space associated to the different categories. Depending on the model complexity, such boundaries can be linear (hyperplanes) or assume more complex (non-linear) shapes. When possible, linear discriminant models are preferred as they have less parameters to tune, require a lower number of training samples and are in general more robust against overfitting. Based on these considerations, the first-ever and still one of the most commonly used discriminant techniques is Linear Discriminant Analysis (LDA), originally proposed by Fisher (1936). It relies on the assumption that the samples of each class are normally distributed around their respective centroids with the same variance/covariance matrix (i.e., the same within-category scatter). Under these assumptions, it is possible to calculate the probability that each sample belongs to a particular class $g$ $p\left(g|\boldsymbol{x}\right)$, as:

$$p\left(g|\boldsymbol{x}\right) = \frac{\pi_g}{C} e^{-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_g)^T \boldsymbol{S}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_g)} \quad (26)$$

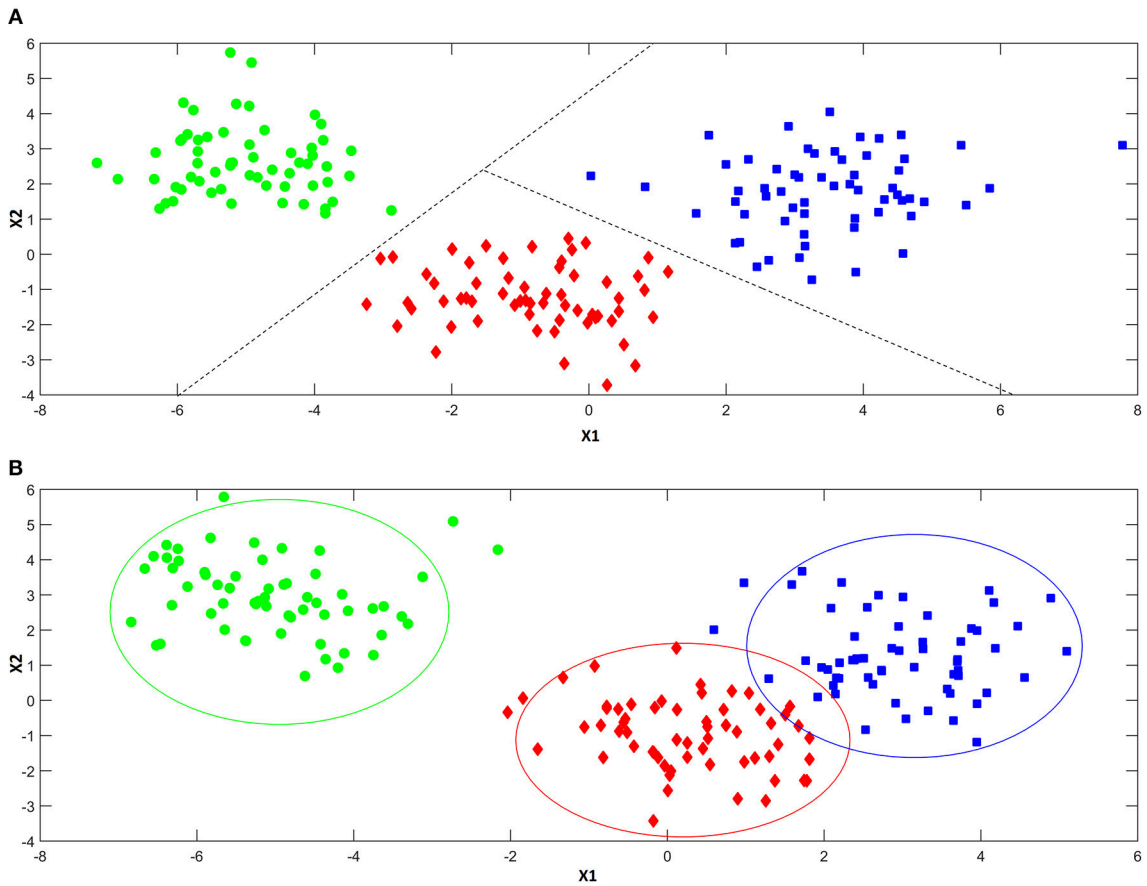



**FIGURE 4 |** Illustration of the difference between discriminant **(A)** and modeling **(B)** classification techniques. Discriminant classification techniques **(A)** divide the available hyperspace into as many regions as the number of the investigated categories (three, in the present example), so that whenever a sample falls in a particular region of space, it is always assigned to the associated class. Modeling techniques **(B)** build a separate model for each one of the categories of interest, so that there can be regions of spaces where more than a class is mapped and others where there is no class at all.

where $\bar{x}_g$ is the centroid of class $g$, $S$ the overall within-class variance/covariance matrix, $\pi_g$ the *prior* probability (i.e., the probability of observing a sample from that category before carrying out any measurement), $C$ is a normalization constant and the argument of the exponential $\left(x-\bar{x}_g\right)^T S^{-1}\left(x-\bar{x}_g\right)$ is defined as the squared Mahalanobis distance of the individual to the center of the category. Classification is then accomplished by assigning the sample to the category, to which it has the highest probability of belonging.

LDA is a well-established technique, which works well also on data for which the normality assumption is not fulfilled but, unfortunately, it can rarely be used on spectroscopic data for the same reasons MLR cannot be utilized for regression (see section Regression): calculation of matrix $S^{-1}$ requires the experimental data matrix to be well-conditioned, which is not the case, when dealing with a high number of correlated variables measured on a limited number of samples. To overcome these limitations, LDA can be applied on the scores of bilinear models used to compress the data (e.g., on principal components), but the most commonly used approach involves a suitable modification of the PLS algorithm which makes it able to deal with classification issues; the resulting method is called partial least squares discriminant analysis (PLS-DA) (Sjöström et al., 1986; Ståle and Wold, 1987; Barker and Rayens, 2003), and it will be briefly described in the following paragraph.

## Partial Least Squares Discriminant Analysis (PLS-DA)

In order for the PLS algorithm to deal with discriminant classification problems, the information about class belonging has to be encoded in a response variable $Y$, which can be then regressed onto the experimental matrix $X$ to provide the predictive model (Sjöström et al., 1986). This is accomplished by defining $Y$ as a "dummy" binary matrix, having as many rows as the number of samples ($N$) and as many columns as the number of classes ($G$). Each row in $Y$ is a vector encoding the information about class belonging of the corresponding sample, whereas each column is associated to a particular class (the first column to class 1, the second to class 2 and so on up to the $G^{th}$). As such, the row vector corresponding to a particular sample will contain all zeros except for the column associated to the class it belongs to, where there will be a one. For instance, in the case of a problem involving three categories, a sample belonging to

Class 2 will be represented by the vector $\boldsymbol{y}_i = [0\ 1\ 0]$. A PLS regression model is then calculated between the experimental data matrix $\boldsymbol{X}$ and the dummy $\boldsymbol{Y}$ [as described in section Partial Least Squares (PLS) Regression], and the matrix of regression coefficients obtained is used to predict the value of the responses on new samples, $\hat{\boldsymbol{Y}}_{new}$. Since the dependent variable is associated to the categorical information, classification of the samples is based on the predicted responses $\hat{\boldsymbol{Y}}_{new}$ which, however, are not binary but real-valued. As a consequence, different approaches have been proposed in the literature to define how to classify samples in PLS-DA based on the values of $\hat{\boldsymbol{Y}}_{new}$. The naivest approach (see e.g., Alsberg et al., 1998) is to assign each sample to the category corresponding to the highest value of the predicted response vector. For instance, if the following predictions were obtained for a particular sample: $\hat{\boldsymbol{y}}_{new,k}$= [0.1 − 0.4 0.8], it would be assigned to Class 3. On the other hand, other strategies have been also suggested, like the application of LDA on $\hat{\boldsymbol{Y}}_{new}$ or on the PLS scores (Nocairi et al., 2004; Indahl et al., 2007), or the use of thresholds based on probability theory (Pérez et al., 2009).

## Class-Modeling Methods

As already stated, class-modeling methods aim at identifying a closed (bound) sub-space, where it is likely to find samples from a particular category, irrespective of whether other classes should also be considered or not. They try to capture the features, which make individuals from the same category similar to one another. Operationally, they define the class space by identifying the "normal" variability which can be expected among samples belonging to that category and, accordingly, introducing a "distance-to-the-model" criterion which accounts for the degree of outlyingness of any new sample. Among the different class-modeling techniques proposed in the literature, soft independent modeling of class analogies (SIMCA) is by far the most commonly used, especially for spectroscopic data, due to its ability of dealing with ill-conditioned experimental data matrices and, therefore, it will be briefly described below (for more details, the reader is referred to Wold, 1976; Wold and Sjöström, 1977, 1987).

### Soft Independent Modeling of Class Analogies (SIMCA)

The main idea behind SIMCA is that the systematic variability characterizing the samples for a particular category can be captured and accurately accounted for by a PCA model of appropriate dimensionality. This model is built by using only the samples from the investigated category:

$$\boldsymbol{X}_g = \boldsymbol{T}_g\boldsymbol{P}_g^T + \boldsymbol{E}_g \tag{27}$$

where the symbols have the same meaning as in Equation (2), and the subscript indicates that the model is calculated by using only the training data from class *g*. The use of PCA to define the similarities among the samples belonging to the category of interest provides also the machinery to assess whether any new sample is likely to come from that class or not through the definition of two statistics normally used for outlier detection, namely $T^2$ and $Q$. As already introduced in section Principal Component Analysis, the former is the squared Mahalanobis distance of a sample to the center of the scores space, indicating how far the individual is from the distribution of the "normal" samples in the space spanned by the significant PCs (Hotelling, 1931), while the latter is the (Euclidean) distance of the sample to its projection onto the PC space, describing how well that individual is fitted by the PCA model (Jackson and Muldholkar, 1979). In the context of SIMCA, once the PCA model of the $g^{th}$ category is calculated according to Equation (27), any specimen to be predicted is projected onto that model and its values of $T^2$ and $Q$ are used to calculate an overall distance to the model $d_{i,g}$ (Yue and Qin, 2001), which constitutes the basis for class acceptance or rejection:

$$d_{i,g} = \sqrt{\left(T_{i,g}^2\right)^2 + \left(Q_{i,g}\right)^2} \tag{28}$$

where the subscript indicates that the $i^{th}$ sample is tested against the model of the $g^{th}$ category. Accordingly, the boundary of the class space is identified by setting a proper threshold to the distance, so that if a sample has a distance to the model lower than the threshold it is accepted by the category and, otherwise, it is rejected.

## Selected Applications of Classification Approaches for Pharmaceutical Analysis

As mentioned before, classification approaches are widely applied in quality controls of pharmaceuticals, in particular to detect counterfeit drugs, as, for instance, it is reported in da Silva Fernandes et al. (2012), where NIR and fluorescence spectroscopy were combined with different classification methods to distinguish among pure and adulterated tablets. In Storme-Paris et al. (2010), a non-destructive approach is proposed to distinguish genuine tablets from counterfeit or recalled (from the market) medicines. In order to achieve this, NIR spectra (directly collected on the tables) are analyzed by SIMCA. Results obtained suggest the validity of this approach; in fact, it allowed highlighting small differences among drugs (e.g., different coating), and it provided an excellent differentiation among genuine and counterfeits products. For the same purpose, namely counterfeit drug detection, NIR spectra were also widely combined with PLS-DA. Only to mention one, de Peinder et al. (2008) demonstrated the validity of this approach to spot counterfeits of a specific cholesterol-lowering medicine. Despite the fact that the authors highlighted the storage conditions sensibly affecting NIR spectra (because of humidity), the PLS-DA model still proved to be robust and provided excellent predictions.

# VALIDATION

Chemometrics relies mainly on the use of empirical models which, given the experimental measurements, should summarize the information of the data, reasonably approximate the system under study, and allow predictions of one or more properties of interest. Bearing this in mind, given the "soft" (i.e., empirical) nature of the models employed, there are many models one could

in principle calculate on the same data and their performances could be influenced by different factors (number of samples and their representativeness, the method itself, the algorithm, and so on) (Brereton et al., 2018). Thus, selecting which model is the most appropriate for the data under investigation and verifying how reliable it is, is of fundamental importance and the chemometric strategies for doing so are collectively referred to as validation (Harshman, 1984; Westad and Marini, 2015). To evaluate the quality of the investigated models, the validation process requires the definition of suitable diagnostics, which could be based on model parameters but more often rely on the calculation of some sort of residuals (i.e., error criteria). In this context, in order to avoid overoptimism or, in general, to obtain estimates which are as unbiased as possible, it is fundamental that the residuals which are used for validation are not generated by the application of the model to the data it has been built on, since in almost all cases, they cannot be considered as representative of the outcomes one would obtain on completely new data. For such reason, a correct validation strategy should involve the estimation of the model error on a dataset different than the one used for calculating the model parameters. This is normally accomplished through the use of an external test set or cross-validation.

The use of a second, completely independent, set of data for evaluating the performances and, consequently, calculating the residuals (test set validation) is the strategy which best mimics how the model will be routinely used, and it is therefore the one to be preferred, whenever possible. On the other hand, cross-validation is based on the repeated resampling of the dataset, into a training and a test sub-sets, so that at each iteration only a part of the original samples is used for model building while the remaining individuals are left out for validation. This procedure is normally repeated up to the moment when each sample has been left out at least once or, anyway, for a prespecified number of iterations. Cross-validation is particularly suited when the number of available samples is small and there is no possibility of building an external test set, but the resulting estimates can be still biased as the calibration and validation sets are never completely independent on one another. In general, it is rather used for model selection (e.g., estimating the optimal number of components) than for the final validation stage.

## OTHER SELECTED APPLICATIONS

In addition to some specific applications described above, in this paragraph additional examples will be presented to further emphasize the usefulness of chemometrics-based spectroscopy for pharmaceutical analysis.

Morris and Forbes (2001) coupled NIR spectroscopy with multivariate calibration for quantifying narasin chloroform-extracted from granulated samples. In another study, Forbes et al. (2001) proposed a transmission NIR spectroscopy method using multivariate regression for the quantification of potency and lipids in monensin fermentation broth.

Ghasemi and Niazi (2007) developed a spectrophotometric method for the direct quantitative determination of captopril in pharmaceutical preparation and biological fluids (human plasma and urine) samples. Since the spectra were recorded at various pHs (from 2.0 to 12.8), different models were tested, including the possibility of a preliminary spectral deconvolution using multi-way approaches. In particular, the use of PLS on the spectra at pH 2.0 allowed to build a calibration curve which resulted in a very good accuracy. Li et al. (2014) used Raman spectroscopy to identify anisodamine counterfeit tablets with 100% predictive accuracy and, at the same time, NIR spectroscopy to discriminate genuine anisodamine tablets from 5 different manufacturing plants. In the latter case, PLS-DA models were found to have 100% recognition and rejection rates. Willett and Rodriguez (2018) implemented a rapid Raman assay for on-site analysis of stockpiled drugs in aqueous solution, which was tested on Tamiflu (oseltamivir phosphate) by using three different portable and handheld Raman instruments. PLS regression models yielded an average error with respect to the reference HPLC values, which was lower than 0.3%. Other examples of application can be found in Forina et al. (1998), Komsta (2012), Hoang et al. (2013), and Lohumi et al. (2017).

## CONCLUSIONS

Chemometrics provide a wealth of techniques for both the exploratory analysis of multivariate data as well as building reliable calibration and classification strategies to predict quantitative and qualitative responses based on the experimental profiles collected on the samples. Coupled to spectroscopic characterization, it represents an indispensable and highly versatile tool for pharmaceutical analysis at all levels.

## AUTHOR CONTRIBUTIONS

AB and FM jointly conceived and designed the paper, and wrote the manuscript. All authors agreed on the content of the paper and approved its submission.

## REFERENCES

Albano, C., Dunn, W., Edlund, U., Johansson, E., Nordén, B., Sjöström, M., et al. (1978). Four levels of pattern recognition. *Anal. Chim. Acta* 103, 429–443. doi: 10.1016/S0003-2670(01)83107-X

Alsberg, B. K., Kell, D. B., and Goodacre, R. (1998). Variable selection in discriminant partial least-squares analysis. *Anal. Chem.* 70, 4126–4133. doi: 10.1021/ac980506o

Alvarenga, L., Ferreira, D., Altekruse, D., Menezes, J. C., and Lochmann, D. (2008). Tablet identification using near-infrared spectroscopy (NIRS) for pharmaceutical quality control. *J. Pharm. Biomed. Anal.* 48, 62–69. doi: 10.1016/j.jpba.2008.05.007

Barker, M., and Rayens, W. (2003). Partial least squares for discrimination. *J. Chemometr.* 17, 166–173. doi: 10.1002/cem.785

Bautista, R. D., Aberásturi, F. J., Jiménez, A. I., and Jiménez, F. (1996). Simultaneous spectrophotometric determination of drugs in pharmaceutical preparations using multiple linear regression and partial least-squares

regression, calibration and prediction methods. *Talanta* 43, 2107–2121. doi: 10.1016/S0039-9140(96)01997-2

Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L., Nescatelli, R., and Marini, F. (2013). "Classification and class-modelling," in *Chemometrics in Food Chemistry*, ed F. Marini (Oxford, UK: Elseivier), 171–232.

Blanco, M., Eustaquio, A., González, J. M., and Serrano, D. (2000). Identification and quantitation assays for intact tablets of two related pharmaceutical preparations by reflectance near-infrared spectroscopy: validation of the procedure. *J. Pharm. Biomed. Sci*. 22, 139–148. doi: 10.1016/S0731-7085(99)00274-5

Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., et al. (2018). Chemometrics in analytical chemistry—part II: modeling, validation, and applications. *Anal. Bioanal. Chem*. 410, 6691–6704. doi: 10.1007/s00216-018-1283-4

Chen, H., Lin, Z., and Tan, C. (2018). Nondestructive discrimination of pharmaceutical preparations using near-infrared spectroscopy and partial least-squares discriminant analysis. *Anal. Lett*. 51, 564–574. doi: 10.1080/00032719.2017.1339070

Cruz Sarraguça, M., and Almeida Lopes, J. (2009). Quality control of pharmaceuticals with NIR: from lab to process line. *Vib. Spectrosc*. 49, 204–210. doi: 10.1016/j.vibspec.2008.07.013

da Silva Fernandes, R., da Costa, F. S., Valderrama, P., Março, P. H., and de Lima, K. M. (2012). Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods. *J. Pharm. Biomed. Anal*. 66, 85–90. doi: 10.1016/j.jpba.2012.03.004

de Peinder, P., Vredenbregt, M. J., Visser, T., and de Kaste, D. (2008). Detection of Lipitor® counterfeits: A comparison of NIR and Raman spectroscopy in combination with chemometrics. *J. Pharm. Biomed. Anal*. 47, 688–694. doi: 10.1016/j.jpba.2008.02.016

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen*. 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Forbes, R. A., Luo, M. Z., and Smith, D. R. (2001). Measurement of potency and lipids in monensin fermentation broth by near-infrared spectroscopy. *J. Pharm. Biomed. Anal*. 25, 239–246. doi: 10.1016/S0731-7085(00)00497-0

Forina, M., Casolino, M. C., and De La Pezuela Martinez, C. (1998). Multivariate calibration: applications to pharmaceutical analysis. *J. Pharm. Biomed. Anal*. 18, 21–33. doi: 10.1016/S0731-7085(98)00153-8

Geladi, P., and Kowalski, B. R. (1986). Partial least squares regression: a tutorial. *Anal. Chim. Acta* 185, 1–17. doi: 10.1016/0003-2670(86)80028-9

Ghasemi, N., and Niazi, A. (2007). Determination of captopril in pharmaceutical preparation and biological fluids using two- and three-way chemometrics methods. *Chin. Chem. Lett*. 18, 427–430. doi: 10.1016/j.cclet.2007.02.019

Harshman, R. (1984). "How can I know if it's real? A catalogue of diagnostics for use with three-mode factor analysis and multidimensional scaling," in *Research Methods for Multimode Data Analysis,* ed H. G. Law, C. W. Snyder Jr, J. Hattie, and R. P. Mc Donald (New York, NY: Praeger), 566–591.

Herkert, T., Prinz, H., and Kovar, K. (2001). One hundred percent online identity check of pharmaceutical products by near-infrared spectroscopy on the packaging line. *Eur. J. Pharm. Biopharm*. 51, 9–16. doi: 10.1016/S0939-6411(00)00126-0

Hoang, V. D., Nhung, N. P., and Aboul-Enein, H. Y. (2013). Recent developments and applications of derivative spectrophotometry in pharmaceutical analysis. *Curr. Pharm. Anal*. 9, 261–277. doi: 10.2174/1573412911309030005

Hotelling, H. (1931). The generalization of Student's ratio. *Ann. Math. Stat*. 2, 360–378. doi: 10.1214/aoms/1177732979

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *Br. J. Stat. Psychol*. 10, 69–79. doi: 10.1111/j.2044-8317.1957.tb00179.x

Indahl, U. G., Martens, H., and Naes, T. (2007). From dummy regression to prior probabilities in PLS-DA. *J. Chemometr*. 21, 529–536. doi: 10.1002/cem.1061

Jackson, J. E., and Muldholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics* 21, 341–349. doi: 10.1080/00401706.1979.10489779

Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Appl. Stat*. 16, 225- 236 doi: 10.2307/2985919

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *J. R. Stat. Soc. C* 31, 300–303. doi: 10.2307/2348005

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York, NY: Springer.

Kendall, M. G. (1957). *A Course in Multivariate Analysis*. London: Hafner.

Kjeldahl, K., and Bro, R. (2010). Some common misunderstandings in chemometrics. *J. Chemometr*. 24, 558–564. doi: 10.1002/cem.1346

Komsta, L. (2012). Chemometrics in pharmaceutical analysis. *J. AOAC Int*. 95, 606–608. doi: 10.5740/jaoacint.SGE_Komsta_intro

Li, L., Zang, H., Li, J., Chen, D., Li, T., and Wang, F. (2014). Identification of anisodamine tablets by Raman and near-infrared spectroscopy with chemometrics. *Spectroch. Acta A* 127, 91–97. doi: 10.1016/j.saa.2014.02.022

Lohumi, S., Kim, M. S., Qin, J., and Cho, B.-K. (2017). Raman imaging from microscopy to macroscopy: quality and safety control of biological materials. *Trends Anal. Chem*. 93, 183–198. doi: 10.1016/j.trac.2017.06.002

Marini, F., Bucci, R., Ginevro, I., and Magrì, A. L. (2009). Coupling of IR measurements and multivariate calibration techniques for the determination of enantiomeric excess in pharmaceutical preparations. *Chemometr. Intell. Lab. Syst*. 97, 52–63. doi: 10.1016/j.chemolab.2008.07.012

Martens, H., and Geladi, P. (2004). "Multivariate Calibration," in *Encyclopedia of Statistical Sciences*, eds S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson (Hoboken, NJ: Wiley), 5177–5189.

Martens, H., and Naes, T. (1991). *Multivariate Calibration*. New York, NY: John Wiley & Sons.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc*. 60, 234–256. doi: 10.1080/01621459.1965.10480787

Mazurek, S., and Szostak, R. (2006a). Quantitative determination of captopril and prednisolone in tablets by FT-Raman spectroscopy. *J. Pharm. Biomed. Anal*. 40, 1225–1230. doi: 10.1016/j.jpba.2005.03.047

Mazurek, S., and Szostak, R. (2006b). Quantitative determination of diclofenac sodium and aminophylline in injection solutions by FT-Raman spectroscopy. *J. Pharm. Biomed. Anal*. 40, 1235–1242. doi: 10.1016/j.jpba.2005.09.019

Morris, D. L. Jr., and Forbes, R. A. (2001). A method for measuring potency of narasin extracts using near-IR spectroscopy. *J. Pharm. Biomed. Anal*. 24, 437–451. doi: 10.1016/S0731-7085(00)00469-6

Myakalwar, A. K., Sreedhar, S., Barman, I., Dingari, N. C., Venugopal Rao, S., Prem Kiran, P., et al. (2011). Laser-induced breakdown spectroscopy-based investigation and classification. *Talanta* 87, 53–59. doi: 10.1016/j.talanta.2011.09.040

Nocairi, H., Qannari, E. M., Vigneau, E., and Bertrand, D. (2004). Discrimination on latent components with respect to patterns. Application to multicollinear data. *Comput. Stat. Data Anal*. 48, 139–147. doi: 10.1016/j.csda.2003.09.008

Pearson, K. (1901). On lines and plans of closes fit to systems of points in space. *Philos. Mag*. 2, 559–572. doi: 10.1080/14786440109462720

Pérez, N. F., Ferré, J., and Boqué, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometr. Intell. Lab. Syst*. 95, 122–128. doi: 10.1016/j.chemolab.2008.09.005

Rigoni, L., Venti, S., Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L., et al. (2014). Quantification of the enantiomeric excess of two APIs by means of near infrared spectroscopy and chemometrics. *Chemometr. Intell. Lab. Syst*. 133, 149–156. doi: 10.1016/j.chemolab.2014.02.004

Rodionova, O. Y., Balyklova, K. S., Titova, A. V., and Pomerantsev, A. L. (2018). Application of NIR spectroscopy and chemometrics for revealing of the 'high quality fakes' among the medicines. *Forensic Chem*. 8, 82–89. doi: 10.1016/j.forc.2018.02.004

Rodionova, O. Y., Houmøller, L. P., Pomerantsev, A. L., Geladi, P., Burger, J., Dorofeyev, V. L., et al. (2005). NIR spectrometry for counterfeit drug detection A feasibility study. *Anal. Chim. Acta* 549, 151–158. doi: 10.1016/j.aca.2005.06.018

Roggo, Y., Edmond, A., Chalus, P., and Ulmschneider, M. (2005). Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms. *Anal. Chim. Acta* 535, 79–87. doi: 10.1016/j.aca.2004.12.037

Sjöström, M., Wold, S., and Söderström, B. (1986). "PLS discriminant plots," in *Pattern Recognition in Practice*, ed E. S. Gelsema, and L. N. Kanal (Amsterdam: Elsevier), 461–470.

Ståle, L., and Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J. Chemometr*. 1, 185–196. doi: 10.1002/cem.1180010306

Storme-Paris, I., Rebiere, H., Matoga, M., Civade, C., Bonnet, P. A., Tissier, M. H., et al. (2010). Challenging Near InfraRed Spectroscopy discriminating ability for counterfeit pharmaceuticals detection. *Anal. Chim. Acta* 658, 163–174. doi: 10.1016/j.aca.2009.11.005

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Westad, F., and Marini, F. (2015). Validation of chemometric models—a tutorial. *Anal. Chim. Acta* 893, 14–24. doi: 10.1016/j.aca.2015.06.056

Willett, D. R., and Rodriguez, J. D. (2018). Quantitative Raman assays for on-site analysis of stockpiled drugs. *Anal. Chim. Acta*. 1044, 131–137. doi: 10.1016/j.aca.2018.08.026

Wold, S. (1976). Pattern Recognition by means of disjoint principal components models. *Pattern Recognit*. 8, 127–139. doi: 10.1016/0031-3203(76)90014-5

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis, Chemometr. *Intell. Lab. Syst*. 2, 37–52. doi: 10.1016/0169-7439(87)80084-9

Wold, S., Martens, H., and Wold, H. (1983). "The multivariate calibration problem in chemistry solved by the PLS methods," in *Matrix Pencils: Proceedings Of A Conference Held at Pite Havsbad, Sweden*, eds A. Ruhe, and B. Kagstrøm (Heidelberg: Springer Verlag), 286–293.

Wold, S., and Sjöström, M. (1977). "SIMCA: a method for analysing chemical data in terms of similarity and analogy" in *Chemometrics, Theory and Application*, ed B.R. Kowalski (Washington, DC: American Chemical Society Symposium Series No. 52), 243–282.

Wold, S., and Sjöström, M. (1987). Comments on a recent evaluation of the SIMCA method. *J. Chemom*. 1, 243–245. doi: 10.1002/cem.1180010406

Xie, Y., Songa, Y., Zhang, Y., and Zhao, B. (2010). Near-infrared spectroscopy quantitative determination of Pefloxacin mesylate concentration in pharmaceuticals by using partial least squares and principal component regression multivariate calibration. *Spectrochim. Acta A* 75, 1535–1539. doi: 10.1016/j.saa.2010.02.012

Yao, J., Shi, Y. Q., Li, Z. R., and Jin, S. H. (2007). Development of a RP-HPLC method for screening potentially counterfeit anti-diabetic drugs. *J. Chromatogr. B* 853, 254–259. doi: 10.1016/j.jchromb.2007.03.022

Yue, H. H., and Qin, S. J. (2001). Reconstruction-based fault identification based on a combined index. *Ind. Eng. Chem. Res*. 40, 4403–4414 doi: 10.1021/ie000141+

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** info@frontiersin.org | +41 21 510 17 00

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership